

Generation and representation of synthetic smart meter data

Tianzhen Hong¹ (✉), Daniel Macumber², Han Li¹, Katherine Fleming², Zhe Wang¹

1. Lawrence Berkeley National Laboratory, Berkeley, California, USA

2. National Renewable Energy Laboratory, Golden, Colorado, USA

Abstract

Advanced energy algorithms running at big-data scale will be necessary to identify, realize, and verify energy savings to meet government and utility goals of building energy efficiency. Any algorithm must be well characterized and validated before it is trusted to run at these scales. Smart meter data from real buildings will ultimately be required for the development, testing, and validation of these energy algorithms and processes. However, for initial development and testing, smart meter data are difficult to work with due to privacy restrictions, noise from unknown sources, data accessibility, and other concerns which can complicate algorithm development and validation. This study describes a new methodology to generate synthetic smart meter data of electricity use in buildings using detailed building energy modeling, which aims to capture the variability and stochastics of real energy use in buildings. The methodology can create datasets tailored to represent specific scenarios with known truth and controllable amounts of synthetic noise. Knowledge of ground truth also allows the development and validation of enhanced processes which leverage building metadata, such as building type or size (floor area), in addition to smart meter data. The methodology described in this paper includes the key influencing factors of real-world building energy use including weather data, occupant-driven loads, building operation and maintenance practices, and special events. Data formats to support workflows leveraging both synthetic meter data and associated metadata are proposed and discussed. Finally, example use cases of the synthetic meter data are described to illustrate potential applications.

Keywords

synthetic data,
smart meter data,
EnergyPlus,
data representation,
building energy modeling,
occupant modeling

Article History

Received: 02 February 2020

Revised: 12 May 2020

Accepted: 13 May 2020

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2020

1 Introduction

Utility demand-side management and energy-efficiency pay-for-performance programs rely on predicted energy savings of retrofit packages as well as analysis of real meter data once the retrofits have been completed. Energy algorithm development is a crucial step in enabling these programs and maximizing their impact. These calculations leverage the expanding smart meter infrastructure to calculate realized energy savings for energy conservation measures (ECMs) at scale rather than relying on costly controlled experiments or other methods. Automated measurement and verification algorithms (aka M&V 2.0), which use smart meter data to estimate gross energy savings of energy efficiency projects, have the potential to greatly reduce the cost of estimating energy savings realization. Detailed building energy model calibration algorithms that use smart meter and audit data

enable reliable prediction of energy savings for specific ECMs and buildings.

Although applying energy algorithms to real-world data is the end goal, adequate real datasets are often difficult to obtain and sometimes unavailable to the algorithm developers. Additionally, algorithm validation is made difficult when using real-world data because the ground truth is not known. For example, accurately calculating an algorithm's non-routine event detection rate is not possible when the non-routine event's occurrence in the data is not known. Another example is inadequately-trained learning algorithms, either due to having no metered data available (i.e., new constructions) or for buildings with constantly changing operation over the training period.

Synthetic data have been used to overcome these challenges in many other fields (Knight 2016; Shrivastava et al. 2017; Marr 2018; Nikolaev et al. 2018; Sarkar 2018;

E-mail: thong@lbl.gov

Tian et al. 2018; Dahmen and Cook 2019; Toole 2019). Large amounts of synthetic data, such as time series hourly or sub-hourly whole-building and end-use electricity and gas simulation output data, as well as accompanying ground truth information, such as ECM labels, can be generated using OpenStudio (Guglielmetti et al. n.d.). OpenStudio is a cross-platform (Windows, Mac, and Linux) collection of software tools that supports whole-building energy modeling using EnergyPlus and advanced daylight analysis using Radiance. OpenStudio is an open-source project including graphical interfaces along with a Software Development Kit (SDK). These synthetic smart-meter data can then be used as inputs to algorithms for algorithm development and validation purposes.

In addition to the synthetic data, high-level building characteristics (also referred to as synthetic metadata) about the models used to generate the time series data can be made available to downstream algorithms. These high-level building characteristics may include information about building type, gross floor area, primary HVAC type, etc. This type of information is often available in real-world datasets such as assessor records but is not widely used in current M&V 2.0 algorithms. Data fusion of high-level building parameters with meter data has the potential to enhance algorithm accuracy and other performance metrics in real-world applications. A combination of synthetic data and synthetic metadata can be used to develop enhanced data fusion algorithms, to identify which pieces of metadata are most useful, and to characterize improvements in algorithm performance. Figure 1 illustrates the current workflow for M&V 2.0 algorithms using synthetic data as well as the proposed enhanced workflow using both synthetic data and synthetic metadata.

We describe the methodology for developing OpenStudio models to generate synthetic data (e.g., time series metered energy use) and its accompanying metadata (e.g., building type, building area, and climate zone). The synthetic data are expected to capture the dynamics and diversity of real

energy use in buildings considering major influencing factors such as weather, occupant behavior, building operation and maintenance, as well as some special events like extreme weather, changes in building operation, or demand response. We also discuss possible representations and formats of the synthetic data and metadata to facilitate data exchange as well as the storage and usage of the data. A follow-up paper will present methodology and results of validating the synthetic smart meter data against real smart meter data.

2 Methodology

Building energy use is strongly influenced by six factors identified in IEA EBC Annex 53 (Yoshino et al. 2017): (1) weather, (2) building envelope, (3) building services and energy systems, (4) building operation and maintenance, (5) occupant activities and behavior, and (6) indoor environmental quality provided. Current simulation practices simplify these factors with static and homogeneous settings or schedules (e.g., the U.S. DOE reference building models), which can not represent the real dynamics and stochastics of energy use in buildings (Yan et al. 2017). Our methodology covers all these six factors and aims to capture their influences on the variability of real energy use in buildings:

- (1) we use various types of weather data in the building energy simulation to generate the synthetic meter data, including the TMY3 and recent years' historical AMYs.
- (2) we use the DOE reference models at various vintages which comply with the minimum requirements of ASHRAE standard 90.1 to represent various efficiency levels and configurations of building envelope and energy systems.
- (3) we define three scenarios (good, average, poor) of building operation and maintenance practices.
- (4) we model occupant behavior, including occupant interactions with building systems and occupant thermal comfort preferences, in detail to capture its complexity using a suite of existing occupant behavior modeling tools.

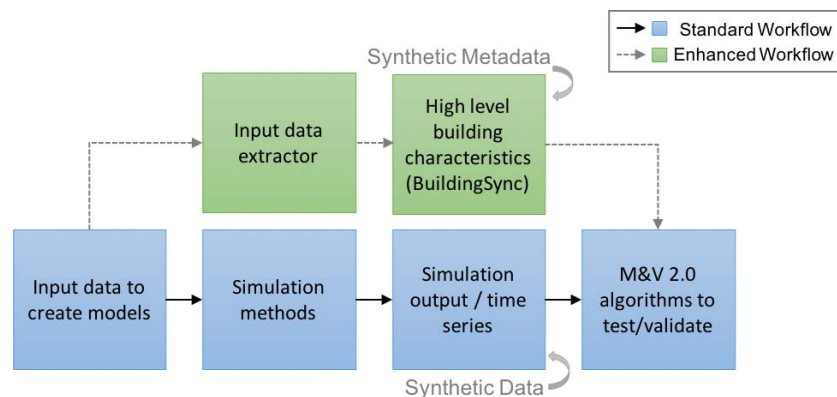


Fig. 1 M&V 2.0 workflows using synthetic data

This section describes the baseline building energy models (i.e., the original DOE reference models) as well as the data and assumptions used to refine the aforementioned influencing factors in said models in order to generate the synthetic meter data.

2.1 Description of the baseline building models

We choose to use the 16 DOE commercial reference buildings (Table 1) in 16 U.S. climates (Table 2) and at five vintages. The baseline scenario is directly from the Commercial Reference Buildings developed by the U.S. Department of Energy (DOE). DOE has provided a complete description

Table 1 16 building types specified in the DOE commercial reference buildings

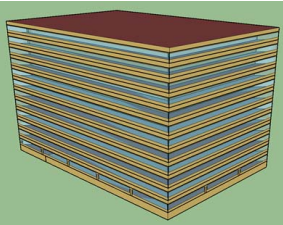
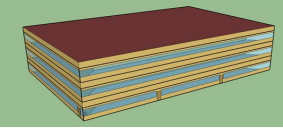
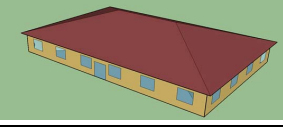
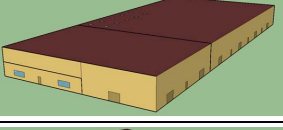
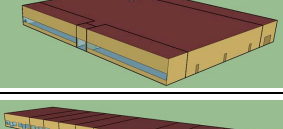
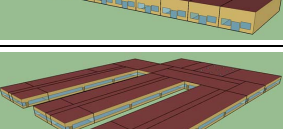
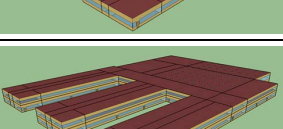
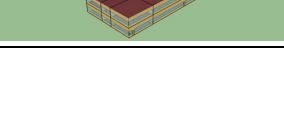
Building type	Floor area (m ²)	Number of floors	Model view
Large office	46,320	12	
Medium office	4,982	3	
Small office	511	1	
Warehouse	4,835	1	
Stand-alone retail	2,319	1	
Strip mall	2,090	1	
Primary school	6,871	1	
Secondary school	19,592	2	

Table 1 16 building types specified in the DOE commercial reference buildings (Continued)

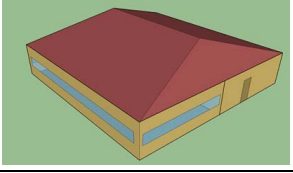
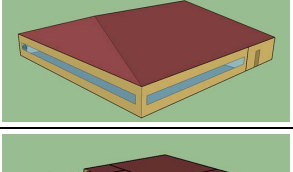
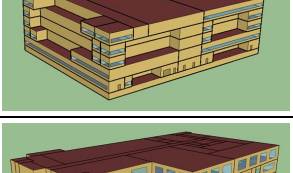
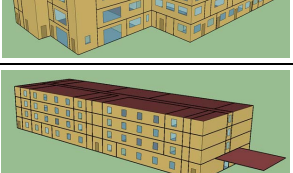
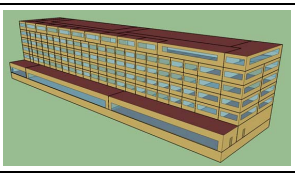
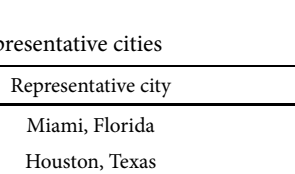
Building type	Floor area (m ²)	Number of floors	Model view
Quick service restaurant	232	1	
Full service restaurant	511	1	
Hospital	22,422	5	
Outpatient health care	3,804	3	
Small hotel	4,013	4	
Large hotel	11,345	6	

Table 2 U.S. climate zones and representative cities

Climate zone	Representative city
1A	Miami, Florida
2A	Houston, Texas
2B	Phoenix, Arizona
3A	Atlanta, Georgia
3B-Coast	Los Angeles, California
3B	Las Vegas, Nevada
3C	San Francisco, California
4A	Baltimore, Maryland
4B	Albuquerque, New Mexico
4C	Seattle, Washington
5A	Chicago, Illinois
5B	Boulder, Colorado
6A	Minneapolis, Minnesota
6B	Helena, Montana
7	Duluth, Minnesota
8	Fairbanks, Alaska

of the whole building information for simulation purposes for those reference buildings (Deru et al. 2011). 16 building types are specified, which represent approximately 70% of the commercial buildings in the U.S. Table 1 lists the major information of the 16 building types.

The five building vintages correspond to the major editions of the national building energy standard ASHRAE 90.1 for 2004, 2007, 2010, 2013, and 2016.

These baseline building models are generated with OpenStudio Measures (Roth et al. 2016) and OpenStudio-Standards Gem (Haves et al. 2017), which is a Ruby library that extends the OpenStudio SDK to implement rules defined in ASHRAE Standards 90.1, 55 and 62.1. The created baseline models are modified through a suite of OpenStudio Measures to implement the detailed models of the aforementioned influencing factors, which are then used to generate the synthetic meter data capturing the variability in real buildings.

2.2 Representation of influencing factors

2.2.1 Building geometry

Building geometry (i.e., floor area, number of stories, floor height, orientation, and aspect ratio) is the basis of creating a building energy model. Building geometry significantly influences heating, cooling, and lighting energy demands, and further affects the type and capacity of the building systems. It is usually the first step to create a geometry model before adding other modeling assumptions such as energy systems, occupancy and operation schedules. Traditionally, DOE commercial reference models (shown in Table 1) are widely used to generate interval load profiles (Field et al. 2010). However, those reference models do not represent the geometry variations in real buildings, which are needed to create a set of synthetic data for an entire building stock (e.g., a district or a city). Therefore, adding variations in building geometry is an important step in the workflow of creating synthetic data. We use OpenStudio-BuildingSync-Gem to create different building geometries starting from the baseline models.

2.2.2 Weather

Weather data (e.g., outdoor air temperature, humidity, pressure, wind speed, solar radiation) is one of the most important inputs for building energy simulation and strongly influences the building energy use (Hong et al. 2013) since it serves as the boundary condition for thermal modeling. In traditional building energy modeling, the input weather data is typically represented as a typical meteorological year (TMY) weather file. The TMY data set is developed by looking at the weather data over a long period in history, typically 30 years. The currently used TMY data set is the

third collection of TMY (TMY3), which was based on weather data for 1020 locations in the USA, derived from a 1976–2005 period of record where available, and a 1991–2005 period of record for all other locations (Wilcox and Marion 2008).

The TMY data set represents the weather condition for a typical year, although it might not necessarily represent the actual weather condition or variations across years. To reflect the yearly weather variation impact on building energy use, we use the actual meteorological year (AMY) weather data. AMY is created from actual hourly data measured by the available ground station for a particular calendar year or for a couple of years (Hong et al. 2013). The AMY weather data can be obtained from several sources, including White Box Technologies, Weather Bank, National Climatic Data Center (NCDC), Weather Source, Weather Analytics, and Meteornorm.

The TMY and AMY weather data are all generated from the weather data recorded in the past. Due to climate change, they are not able to reflect weather dynamics in future. To address this issue, some researchers proposed different approaches to generate future weather data based on different scenarios (Dickinson 2016), which are also used in this workflow.

2.2.3 Occupant behavior

Occupant behavior (Yan et al. 2017) has strong influences on occupant-driven loads, including HVAC, lighting, and appliances. Energy-related occupant behavior is a key factor influencing building performance (Hong et al. 2017; D'Oca et al. 2018). Depending on the building type, climate, and degree of automation in operation and controls, occupants may increase or decrease energy use by a factor of up to three for residential buildings (Andersen 2012), and increase by up to 80% or reduce by up to 50% for single-occupancy offices (Hong and Lin 2013). Occupant behavior in buildings refers to (1) occupant presence in spaces and movement between spaces, (2) occupant interactions with building systems, and (3) occupant adaptations (e.g., changing clothing, having hot/cold drinks). Occupant actions such as adjusting a thermostat for comfort, switching lights on/off, using appliances, opening/closing windows, pulling window blinds up/down, and moving between spaces can have a significant impact on energy use and occupant comfort in buildings.

Unlike the predetermined and fixed occupancy schedules, the real occupant count at space/room level varies randomly, but the distribution of occupancy variations is known. We first develop and implement a more realistic occupant schedule to simulate this behavior. Additionally, the lighting and Miscellaneous Electric Loads (MELs) schedules are related to occupant schedules, for instance office workers might turn off their lights and monitors when they leave

their offices. The dependence of lighting and MELs schedules on occupancy is implemented as well. A more accurate simulation of occupancy, lighting and MELs would not only directly influence the end use load shapes but also enable a more accurate input of internal heat gains, which improves the accuracy of HVAC energy consumption simulation. Additionally, we also consider the occupants' diversified thermal comfort needs and their impact on HVAC energy consumption.

Occupants have more freedom to interact with building equipment and systems in residential and office buildings. Separately, the authors have developed detailed occupant behavior models (Deme Belafi et al. 2019) which can be applied to small, medium, and large office models. These occupant models, or others, can be used in this methodology.

2.2.3.1 Occupant schedules

An agent-based stochastic occupancy model (Chen et al. 2018) will be utilized to simulate occupants' presence and movement in buildings, which generates occupant schedules at the room level and the whole building level. The model is able to capture the spatial and temporal occupancy diversity and stochasticity (Luo et al. 2017). Each occupant and each space in the building are explicitly simulated as an agent with their profiles of stochastic behaviors. The occupancy states are simulated with three types of models: (1) the status transition events (e.g., first arrival in office) simulated with probability distribution models, (2) the random moving events (e.g., from one office to another) simulated with a homogeneous Markov chain model, and (3) the meeting events simulated with a stochastic model. The Occupancy Simulator generates a different set of occupant schedules for a building in each simulation using a random seed. To produce a repeatable set of schedules, Occupancy Simulator can use a fixed seed.

2.2.3.2 Occupants' thermal comfort preference

It has been realized that individuals have different thermal

comfort needs and varying preferred temperature (Wang et al. 2018). Therefore, in buildings where the temperature setpoint is individually adjustable, the temperature setpoint might vary by office-space since inhabitants have varying thermal comfort needs. To take the different thermal demands into consideration, we apply a stochastic temperature setpoint model by assuming that the temperature setpoint follows the normal distribution in the large population (Wang and Hong 2020). To infer the parameters of the normal distribution, we utilized the ASHRAE Global Thermal Comfort Database I (de Dear 1998) and Database II (Földvary Licina et al. 2018), which in total contain 103,846 observations. Focusing on the context of U.S. office buildings, we selected 11,600 data points from the original database. These data were collected in nine U.S. cities: Berkeley, San Francisco, Alameda, Philadelphia, San Ramon, Palo Alto, Walnut Creek, Grand Rapids, Auburn, and the State of Texas (no specific cities were mentioned). To infer the parameters of the normal distribution, Bayesian Inference was utilized. The distribution of occupants' preferred air or operative temperature is calculated, as shown in Table 3, where we also compared the recommended temperature setpoint from the data-driven approach with the DOE Reference Building models.

2.2.4 Lighting

The key assumption for the lighting model is that lighting will be turned off with a 15-min delay once a space is unoccupied. According to the 2012 CBECS, 45% of U.S. commercial buildings have adopted occupancy-detection lighting control. With more and more buildings taking measures to curtail lighting energy consumption, occupancy-based lighting control is expected to be increasingly popular. The occupancy-detection lighting control is always deployed with a time delay to avoid frequent lighting on-off switches. Different buildings might adopt different time delays, varying from 5 to 30 minutes (Guo et al. 2010; Fernandes et al. 2014; de Bakker et al. 2017) with a typical time delay of 15 minutes

Table 3 Inferred thermal neutral air and operative temperatures

		Mean	Standard deviation	5% estimate (recommended lower boundary)	95% estimate (recommended upper boundary)
Air temperature (°C)	Whole period	23.11	1.72	20.37	25.95
	Cooling	23.72	1.19	21.83	25.61
	Heating	22.81	1.87	19.84	25.78
Operative temperature (°C)	Whole period	23.15	1.38	20.96	25.34
	Cooling	23.62	1.09	21.89	25.35
	Heating	22.69	1.36	20.53	24.85
DOE reference building set-point temperature (°C)	Cooling	21.00	/	/	/
	Heating	24.00	/	/	/

(de Bakker et al. 2017). Even for those buildings without occupancy-based lighting control, turning off the light when leaving the space still might happen manually.

2.2.5 MELS (aka plug-loads)

Existing literature, shown in Table 4, has revealed that plug-loads have a strong linear correlation with occupant count. Additionally, field measurement results shown in Figure 2 confirmed the linear relation between MELS and occupant count.

Based on the literature review and the field measurements,

we applied Eq. (1) to model the MELS. Variable “ a ” denotes the base MELS load, referring to those devices that are not turned off during non-working hours (for example, the WiFi routers and desktops that are not shut off). Variable “ b ” denotes the additional MELS load that is controlled and associated with the occupant count.

$$\text{MELS} = a + b \times \text{occ} \quad (1)$$

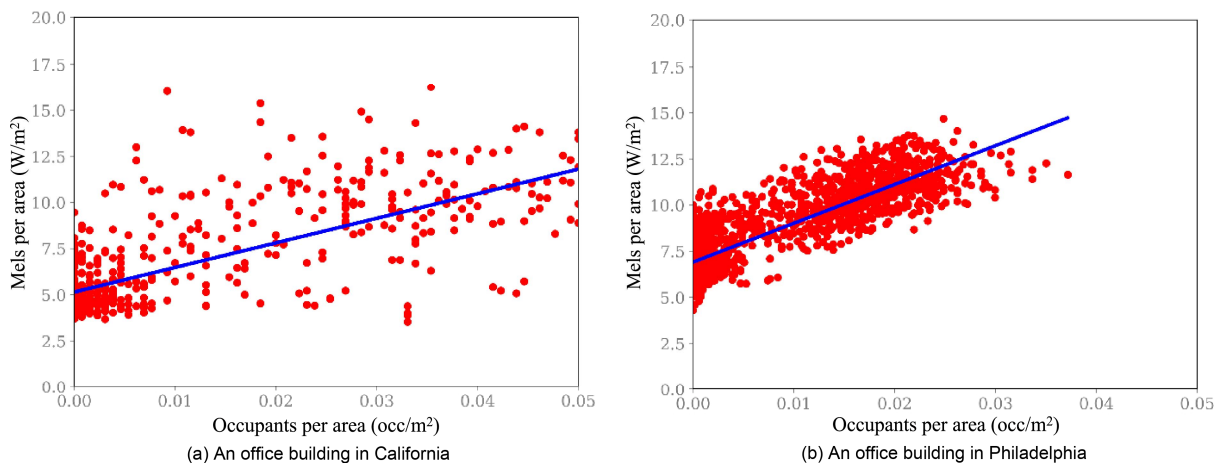
Combining the existing literature and the data collected from real office buildings, we found:

- The base load of MELS (“ a ” in Eq. (1)) is in the range of

Table 4 Existing literature on linear relation between MELS and occupant count

Research		Buildings			Regression	
Literature	Method	Location	Primary usage	Floor area (m ²)	Relation (kW)	R ²
Kim et al. 2017; Kim and Srebric 2017	Field measurement	Philadelphia, PA, USA	Office	6140	$22 + 0.2\text{occ}$	0.74
Mahdavi et al. 2016	Field measurement	Vienna, Austria	1 open plan office + 3 private office	113	Plug load fraction = 10% + occ fraction \times 0.52 (private devices only)	0.70
Martani et al. 2012	Field measurement	Cambridge, MA, USA	Office classroom combination		$9.3 + 0.22\text{WiFi}^1$	0.69
Gunay et al. 2016	Field measurement	Ottawa, Canada	10 office rooms	15 m ² for each office	Occupied: $(0.025 \text{ to } 0.273)\text{occ}$ During night: 0.9 Weekends: 0.5	
Masoso and Grobler 2010	Field measurement	South Africa	5 office buildings		Occupied: 36 W/m ² ; Off-hour: 18 W/m ²	
Dunn and Knight 2005	Survey (30 offices)	Cardiff, UK	Office		$(0.12 \text{ to } 0.23)\text{occ}$	

¹ WiFi connection count can be considered as a proxy variable of occupant count



	An office building in California (CA)	An office building in Pennsylvania (PA)
Base load (W/m ²)	5.1	6.9
Slope (W/occ)	133.1	210.2
R-squared	0.58	0.75

(c) Regression results

Fig. 2 Results from field measurements of real buildings

- 50 – 150 W per peak occupant count;
- The slope of MELs (“*b*” in Eq. (1)) is in the range of 110 – 220 W per occupant count; and
- The base MELs load accounts for around 30% of the peak MELs load.

Finally, we chose a value of 60 for “*a*” and 140 for “*b*” for use in the OpenStudio-occupant-variability GEM to ensure the generated peak MELs load in the models is consistent with the values recommended by the ASHRAE Standard 90.1 or used in the DOE Reference Building models.

2.2.6 Ventilation air

Demand control ventilation (DCV) refers to the ventilation method that supplies fresh air volume based on the indoor occupant count. As a substantial amount of energy is used to heat or cool the fresh air, DCV is an efficient method of reducing building energy consumption (Chao and Hu 2004). According to CBECS 2012, 9.9% of large offices, 6.6% of medium offices and 1.9% of small offices have adopted DCV, as shown in Table 5. For those buildings that have not installed DCV, DCV can still be performed manually rather than automatically. Additionally, due to the energy-saving potential of DCV, it is expected that it will be more widely utilized in the future.

Table 5 Office buildings with DCV according to CBECS 2012

Office type	# buildings (thousand)	# buildings with DCV (thousand)	Proportion of buildings with DCV
Large	5.7	0.6	9.9%
Medium	95.6	6.3	6.6%
Small	911.0	17.0	1.9%

2.2.7 Operation and maintenance practice

Building operation & maintenance (O&M) is another influential factor to be considered in this study. Mathew et al. (2018) defined good, normal and poor practices of building O&M in terms of lighting control, plug load control,

plug load intensity, HVAC schedules, and economizer controls. They found that O&M practice has huge impacts on building energy use. Based on their study, we selected the following assumptions to define two levels (good and poor) of O&M practices, as shown in Table 6.

2.2.8 Special events

Building energy use and load shapes can be affected by special (non-routine) events, e.g., changes in building operation, demand response, extreme weather, wildfire, hurricane, and faulty operations due to failure of sensors, actuators, or equipment. Changes in building operation, such as increased operating hours or tenant change over, often occur in tandem with building upgrades that implement ECMs and can confound savings estimates. Demand response (DR) is a set of time-dependent program activities and tariffs that seek to reduce electricity use or shift usage to another time period. There are multiple ways for the building operator to respond to a DR signal to reduce utility bills. Motegi et al. (2007) summarized some widely used measures in response to a DR signal based on field tests in 28 non-residential buildings, and found that HVAC systems can be a good source for DR savings for three reasons. First, HVAC systems consume a substantial proportion of electricity in buildings. Second, building thermal storage allows for building thermal load shifting. Third, HVAC systems are usually highly automated. In addition to the HVAC systems, lighting, and miscellaneous equipment also provide chances for DR responses. In this project, we will consider the following DR measures based on the literature review, as listed in Table 7.

Wildfire is another special event to be considered, which has become increasingly frequent and severe in recent years. The extreme weather events have caused significant economic damages in recent years (Ranson et al. 2016; Thomas et al. 2016). In 2018, the wildfires in California led to over 24 billion US dollars of damage (Bartz 2019). According to the Fourth National Climate Assessment, there is an expected 30 percent increase in the annual area burned from wildfires by 2060 (Wehner et al. 2017). Wildfire

Table 6 Definition of good and poor O&M practices (adapted from Mathew et al. 2018)

Factors	Good practice	Average practice (used in DOE reference building models)	Poor practice
Occupant density	400 sf/person	200 sf/person	130 sf/person
HVAC schedule	2 hrs before occupancy schedule to turn on HVAC	2 hrs before occupancy schedule to turn on HVAC	Fixed schedule between 6 AM and 8 PM
Supply air temperature reset	Reset base on warmest zones	Reset based on a stepwise function of outdoor air temperature	Constant supply air temperature
VAV box minimum flow settings	15% of the design flow rate	30% of the design flow rate	50% of the design flow rate
Economizer controls	Enthalpy based on ASHRAE 90.1	Dry-bulb temperature based on ASHRAE 90.1	None or broken

Table 7 Demand response measures

System	DR measures
HVAC	Increase/decrease the indoor air temperature setpoint by 4 °F (2.2 °C)
Lighting	Continuous dimming at the office area & zone switching at hallway (reduce the lighting electricity consumption by 33%)
	Continuous dimming at the office area & zone switching at hallway (reduce the lighting electricity consumption by 50%)
	Zone switching at daylit area turn off the artificial lighting if the natural daylight can provide enough lighting

influences the building operation and end-use electricity load shape because it leads to air pollution locally. To respond to the poor outdoor air quality, building occupants tend to close the windows and doors and building operators reduce the fresh air volume to the minimum or fully close outdoor air dampers.

2.3 Synthetic meter data post-processing

Smart meter data in real buildings are subject to various factors which can cause poor data quality. One of the most common issues is sensor and meter drift. Sensors and meters may have biased readings due to poor maintenance or lack of calibration. In addition, equipment malfunction, communication failures, and extreme weather conditions may lead to missing values. Therefore, the generated synthetic meter data can be further manipulated as needed to represent the data quality issues usually seen in real building data such as by adding noise to the data and randomly removing some data in some periods.

2.4 Workflow

Figure 3 illustrates the key steps to generate synthetic meter data and synthetic metadata. The workflow starts with the high-level input of building type, climate zone, vintage, and floor area. It then splits into a synthetic meter data branch and a synthetic metadata branch.

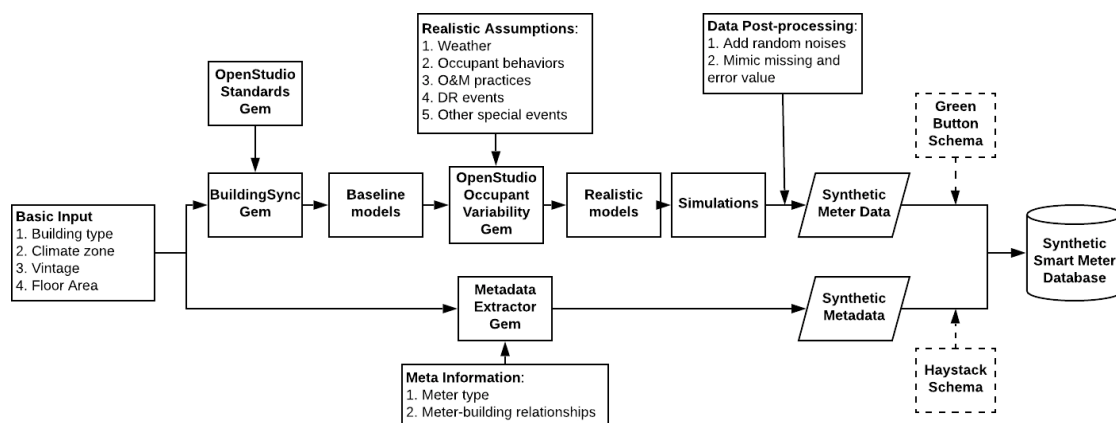
In the meter data branch, the high-level information

is read by the OpenStudio Standards Gem, which creates the baseline models following rules defined in ASHRAE Standards 90.1, 55, and 62.1. Then a series of improved assumptions, including weather, building geometry, occupant behaviors, operation and maintenance practices, demand response events, and other special events are added to the baseline models to create realistic models. This is achieved with OpenStudio Measures, which is included in the OpenStudio-occupant-variability Gem (Li and Macumber 2019).

Next, simulations with the realistic models are conducted to derive the raw synthetic meter data, which are then post-processed, if necessary, by adding random noise and dropping certain values to emulate the real smart meter data. Finally, the synthetic meter data are saved in a database with a common data schema (e.g., Green Button). In the metadata branch, the high-level information can be specified or extracted in a standardized way. Other information such as the meter type, the meter-building relationship are also added to create the synthetic metadata. Similar to the synthetic meter data, the metadata can be saved in the database with a common data schema (e.g., Haystack).

3 Representation of the synthetic meter data and metadata

To complement the synthetic time series meter data, descriptive or input data of the models used to generate said meter data are crucial to maximizing the usefulness of

**Fig. 3** Workflow to generate the synthetic data

synthetic data in developing energy algorithms. We refer to these input data here as metadata.

The metadata include various types of information, such as:

- Building characteristics data—non-changing data about the building, such as building type, square footage, number of floors, window-to-wall ratio, etc.
- Hours of operation/occupancy schedules—the hours that a building is typically occupied over the days of the week, including special circumstances such as holidays.
- Weather data—the weather information, either high-level such as climate zone, or detailed such as the outdoor air temperature and humidity level, used when generating the smart-meter data. Weather data in EnergyPlus format epw is used.
- Special or non-routine events (NREs)—the ground truth information about special event periods that occurred during the synthetic data generation and when they occurred during the generation. These are important when validating that algorithms correctly identify non-routine adjustment periods and atypical operations. An effort by Efficiency Valuation Organization (EVO) to create an International Performance Measurement and Verification Protocol application guide to address non-routine events and adjustments has recently been announced¹. This effort will provide more rigorous descriptions and calculations of the effect of NREs, and is worth further investigation in future studies.
- Smart-meter metadata—information about the smart meters themselves, such as frequency of measurement recording, and what end uses (whole building or lighting, HVAC, DHW, MELs) are connected to the meter.
- Simulation metadata—information about the simulation engine (e.g., OpenStudio, EnergyPlus) and parameters (e.g., number of timesteps per hour, simulation begin and end dates).

3.1 Synthetic meter data

The first step towards a representation of synthetic data is to develop a schema that details the structure and types of the data so that it can be reliably consumed by various applications. Schemas are implemented in one or more formats. Examples of data formats include JSON, XML, and CSV.

Time series data such as synthetic smart-meter data can be represented by a list of data columns, the first of which is a timestamp. Each row then represents data at a specific

timestamp. The format for such data is often CSV.

Green Button is a standardized XML schema used by utilities to represent energy usage data. Customers can get access to their detailed energy data simply by clicking a literal green button on many utility websites. Over 50 utilities have signed agreements to participate in the Green Button initiative, including PG&E and Southern California Edison. This format is an alternative to the common CSV format and can be used to represent the synthetic smart-meter data. It would not, however, be able to accommodate the synthetic metadata.

3.2 Synthetic metadata

Just as in the case of the synthetic data themselves, a schema must be developed to represent the synthetic metadata generated, and a format must be selected to implement the schema.

JSON is a common, lightweight data-exchange option that is easy both for people to understand and for computers to digest, and would be a good format for metadata. An example of how metadata can be represented in JSON format is shown in Figure 4.

The selection of a format to represent synthetic metadata should take into consideration currently available formats used in the building space. Standardizing on a format that is similar or identical to an existing format would be advantageous; in addition to the benefit of reusing a format that people are already familiar with.

Several existing data formats that are being considered to represent synthetic metadata are summarized in Table 8 and described below.

BuildingSync

BuildingSync is a standard XML schema for energy audit data. Its purpose is to standardize energy audit data collection to streamline the process and facilitate data exchange between a variety of software tools and databases in the energy audit space. Full schema specifications and data dictionary can be found at <https://buildingsync.net/schema/>.

```
{
  "building": {
    "square_footage": 50000,
    "number_of_stories": 3,
    "building_type": "small office",
    "climate_zone": "5B",
    "weather_file": "USA_CO_Denver.Intl.AP.725650_TMY3.epw",
    "number_of_occupants": "200",
    "hours_of_operation": "M-F 8am - 5pm"
  },
  "simulation": {
    "engine": "EnergyPlus",
    "timesteps_per_hour": 4
  }
}
```

Fig. 4 Example of metadata in JSON format

¹ <https://evo-world.org/en/news-media/evo-news/1137-efficiency-valuation-organization-to-create-an-ipmvp-application-guide-on-non-routine-events-nre-and-non-routine-adjustments-nra>

Table 8 Summary of data schema in the building smart meter data field

Schema	What is it?	Data format	Potential usage in synthetic data
BuildingSync (BuildingSync n.d.)	BuildingSync is a schema for energy audit data which allows data to be more easily aggregated, compared, and exchanged between different databases and software tools.	XML	To represent synthetic metadata
Project Haystack (Project Haystack n.d.)	Project Haystack is a tagging system for describing building assets using semi-structured sets of tags which aims to allow semantic understanding across the IoT industry.	JSON or CSV	To represent synthetic metadata
BRICK (BrickSchema n.d.)	Brick is a data schema which standardizes semantic descriptions of the physical, logical and virtual assets in buildings and the relationships between them.	RDF	To represent synthetic metadata
Building Energy Data Exchange Specification-BEDES (BEDES n.d.)	BEDES is a dictionary of terms, definitions, and field formats for standardization in terminology and vernacular for quantities including building characteristics, energy transactions, and Internet of Things (IoT).	Text	To standardize the terms defined in the synthetic metadata and synthetic meter data

Project Haystack

Project Haystack is an open-source initiative aimed at standardizing the descriptors used to describe or “tag” metadata. Standard taxonomies include units, energy metering, and various HVAC equipment. Haystack has several data exchange formats, including JSON and CSV. Haystack can be used to standardize units and energy meter descriptors. Figure 5 contains a diagram of the descriptor relationship between an electric meter and its submeter.

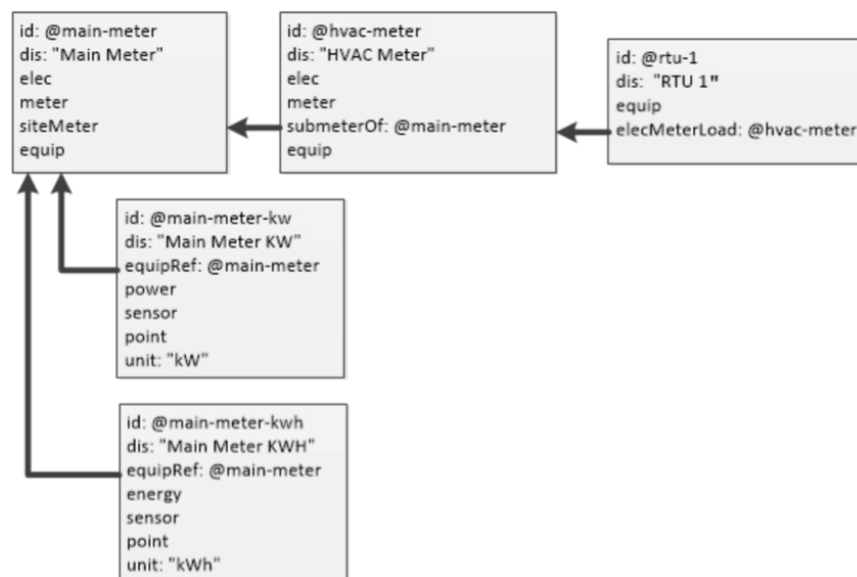
BRICK

The BRICK schema is an open-source semantic representation of building metadata. It is an ontology written in the Resource Description Framework language, a format that represents web resources and their metadata by a directed graph of nodes. Since RDF is meant to be processed by

applications and not humans, the graphs can be written in a text-based syntax called Turtle. BRICK can also make use of all the standard RDF tools for storage, querying, and visualization. BRICK uses RDF vocabularies of tags to represent building components and subsystems. It provides relationships for connecting the components into a directed graph representation of a building.

BEDES

Although not a schema in itself, the Building Energy Data Exchange Specification is a dictionary of terms used in the building energy space to provide standardized definitions. Using BEDES-compliant terms in the synthetic smart-meter metadata schema will facilitate the identification and exchange of building characteristics and energy usage information with other BEDES-compliant applications and formats, such as Green Button, BuildingSync, and SEED.

**Fig. 5** Diagram of an electric meter and its submeter

3.3 Synthetic data storage and retrieval

Just like in the case of real data, there is a need to effectively store and query synthetic data and metadata. Synthetic data can be stored in time series files (e.g., CSV format), while metadata can be stored in a database with links to the associated time series data files. The type of database is not so critical in this application, as the amount of data is not extreme. However, adopting a database that supports JSON format (such as MongoDB or PostgreSQL) can make data transfer easier.

The Standard Energy Efficiency Data (SEED) Platform is a standardized data platform for managing building performance data from a variety of sources. The SEED platform can store the data and metadata. The SEED platform supports the BuildingSync format, which, if used to represent the synthetic metadata, would facilitate storage into the SEED platform. Many cities are using SEED to store energy audit data which can then be retrieved in BuildingSync format. This would provide a central storage location for both synthetic metadata and real metadata obtained from energy audits.

SEED and other applications are in the process of implementing Unique Building Identification (UBID) (Wang et al. 2019) in their platforms. UBID is a standardized framework that facilitates building data matching across different data sources into a single location as well as data exchange. The unique building identifier is based on the Google Open Location Code (OLC) and is generated from the spatial dimensions of a building. The use of this identifier will enable data exchange and collation.

4 Workflow demonstration

A simulation case study is conducted to demonstrate the workflow to generate synthetic meter data. This section discusses the workflow to generate the synthetic meter data using a baseline model and a model considering influencing factors aforementioned.

The baseline model we used is the DOE prototype detailed medium-sized office building model created with

OpenStudio Standards Gem. The building is a three-story rectangular building with a total floor area of 4982 m² located in U.S. climate zone 4A. The detailed model has the same geometry as the original prototype medium-sized model, which can be found in Table 1. However, the detailed version of the models have more sophisticated and realistic space and thermal zone configurations. Figure 6 shows the comparison of the space types between the original and detailed medium-sized office buildings.

In Figure 6, each color represents a unique space type. It can be seen that the original model has all of its spaces represented as type “office”, whereas the detailed model has multiple space types including office, conference, mechanical rooms, storage, restrooms, corridors, and elevator. The heterogeneous space types of the detailed model allow more realistic stochastic occupancy simulations since occupants have different movement behaviors and indoor environmental preferences in different spaces.

A set of influencing factors including stochastic occupancy schedules, lighting schedules, MELs schedules, thermostat setpoint schedules, and demand-controlled ventilation are applied to the model using OpenStudio measures which are included in OpenStudio-occupant-variability Gem. Figure 7 illustrates the number of people in a big office space with the original fixed occupant schedule and the stochastic occupant schedule. It can be seen that the baseline has the same occupant schedules for all weekdays, while the updated model introduces variations to people count at the timestep level.

Figure 8 and Figure 9 show the heatmap of people count in an office space for the baseline model and the updated model, respectively. The updated model introduces dynamics in terms of varied arriving time, lunch break time, and leaving time. It can be seen that the peak number of people may occur at different times of the day.

In addition to occupancy schedule, variabilities are introduced to the settings of lighting, MELs, thermostat setpoints, and mechanical ventilation. As discussed in Section 2.2, those variabilities are correlated with the occupancy schedule.

With the stochastics in occupancy, lighting, MELs,

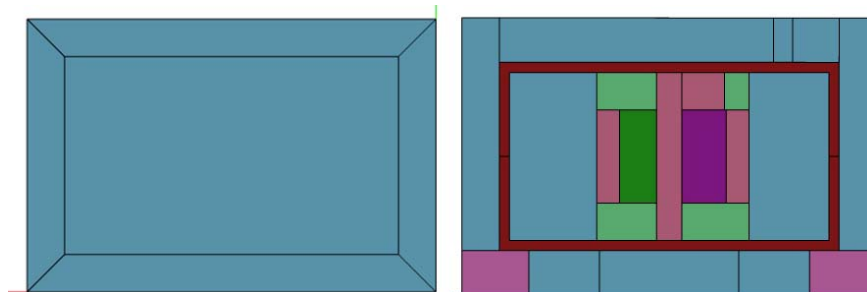


Fig. 6 Space types comparison between the original and detailed DOE prototype medium-sized office building

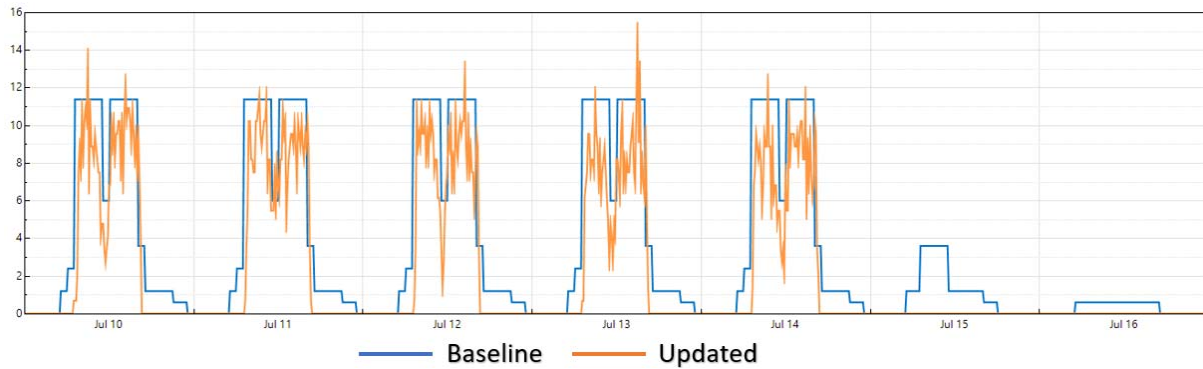


Fig. 7 Number of people in an office space with fixed and stochastic occupancy schedules

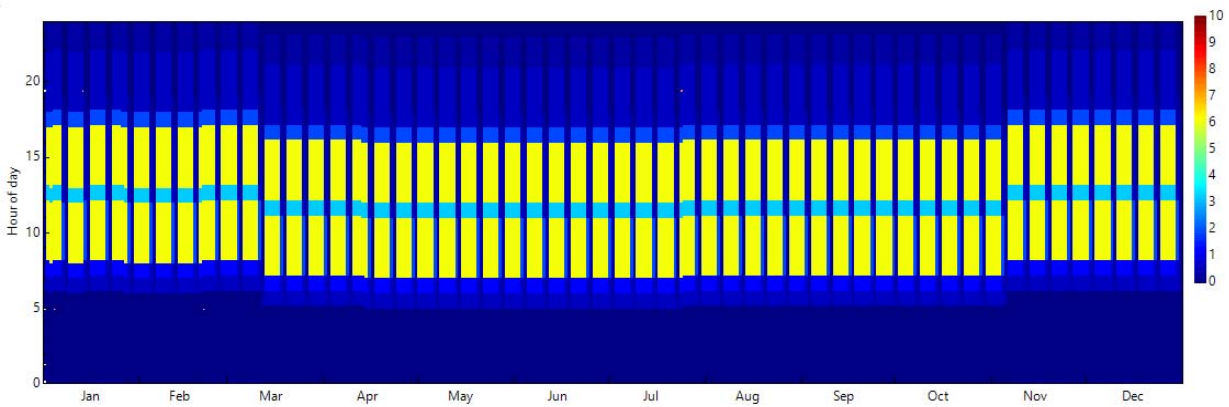


Fig. 8 Heatmap of people count in an office space with fixed occupancy schedules

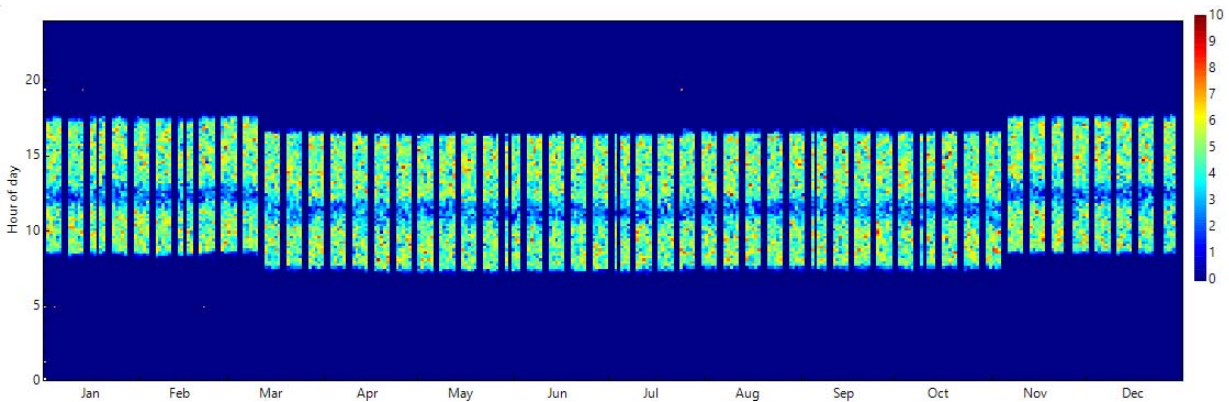


Fig. 9 Heatmap of people count in an office space with stochastic occupancy schedules

mechanical ventilation, and zone air temperature setpoint introduced, the whole-building energy consumption is expected to vary between the baseline model and the updated model. Figure 10 shows the hourly whole-building electricity consumption in a whole year span. The electricity consumption trends of the baseline model and the updated model are similar—the building consumes more electricity in winter and summer months when there are higher heating and cooling demands than in swing seasons.

However, after zooming into the daily level, the difference between the two models becomes obvious. Figure 11 and

Figure 12 indicate the whole-building electricity consumption of the baseline and updated model during an example winter day and an example summer day, respectively. It can be seen that electricity consumption of the baseline model is relatively stable during the working hours in both examples, while the electricity consumption of the updated model vacillates from time to time, due to the stochastic occupant movements and related changes to the schedules of lighting, MELs, and ventilation rate. Another finding is that the stochastic occupancy schedule has different levels of impacts on the whole building electricity consumption in winter

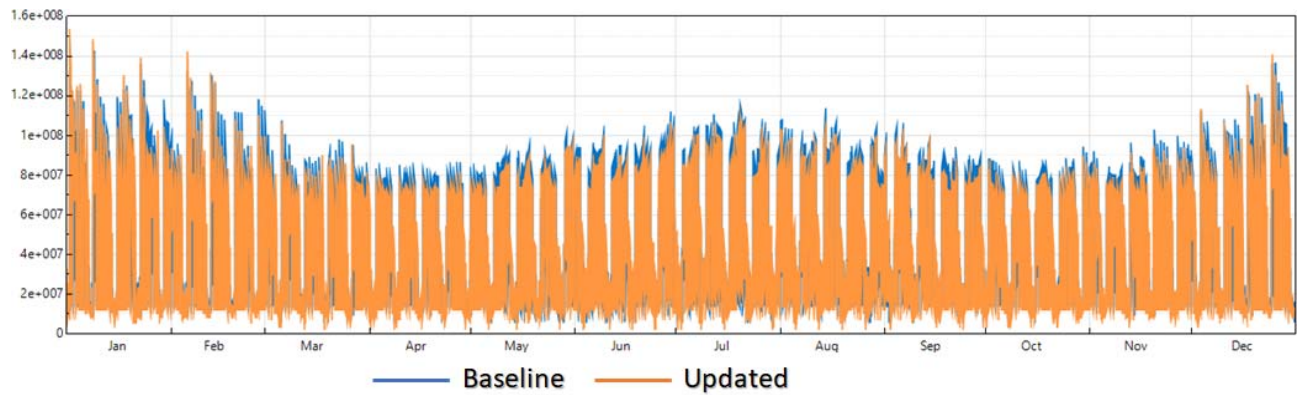


Fig. 10 Whole-building electricity consumption time series in a year

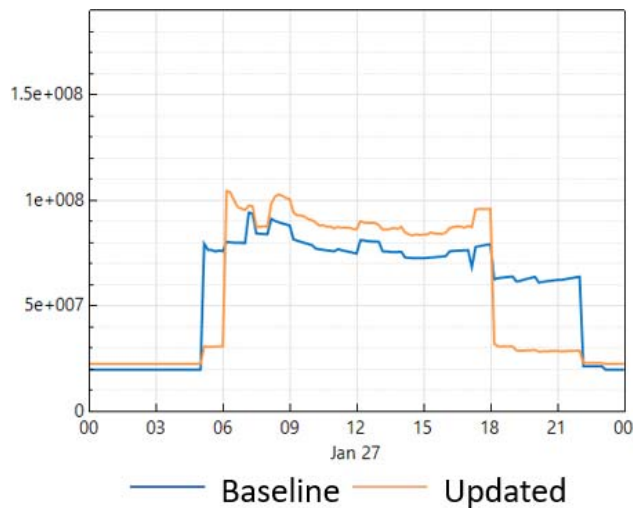


Fig. 11 Whole-building electricity consumption on an example winter day

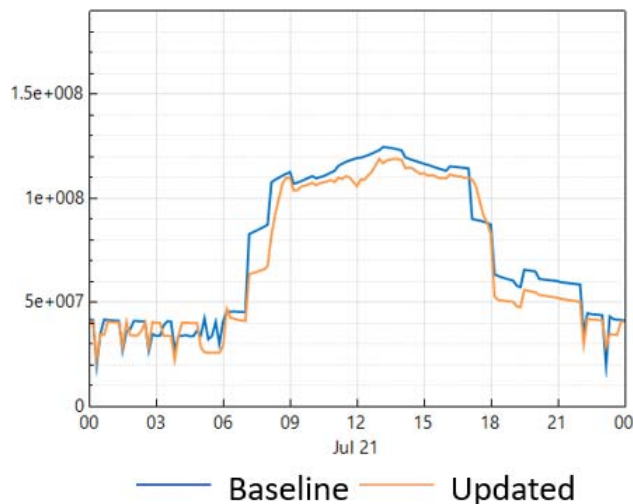


Fig. 12 Whole-building electricity consumption on an example summer day

and summer. For example, in Figure 11, there is a late increase (6 AM) and early decrease (6 PM) of the electricity consumption in the updated comparing with the baseline

model, while in Figure 12, the baseline and updated models show similar times of the increase and decrease of electricity consumption. The reason behind the seasonal difference is that electric lighting and equipment consumption related to occupants account for larger portion of the whole-building consumption in winter than in summer. It should be noted that the occupant arrival and departure time assumptions can be modified when generating stochastic occupancy schedules using the workflow developed in this study.

5 Discussion

5.1 Use cases of the synthetic data

The synthetic time series meter data can be used in various applications of utility energy efficiency programs as well as in the building life-cycle to improve energy efficiency.

A major use case is energy algorithm testing and validation. Algorithms in the energy space predict savings and identify and evaluate opportunities for energy retrofit and operational improvements in buildings. The main class of energy algorithms is called Automated Measurement and Verification, or M&V 2.0 algorithms. These algorithms calculate energy savings based on meter data collected before and after an improvement has been made in a building. An example M&V 2.0 algorithm is the CalTRACK 2.0 methods for computing Normalized Metered Energy Consumption (NMEC), which is a standardized calculation that can be used to determine site-level energy savings from installed efficiency projects.

While M&V 2.0 algorithms are implemented to be used on real-world data, these algorithms should be tested and validated on a clean and complete dataset. Synthetic data can provide these clean datasets while also providing the “ground truth” so that the prediction accuracy of the algorithm can be calculated and analyzed. Once the algorithms have been tested and validated using clean synthetic data, noise can be increasingly added to explore how the algorithm’s

performance is sensitive to noise. Because the ground truth is known for synthetic data, accuracy can be explored and error metrics developed. This provides some understanding of the algorithm's characteristics and accuracy when using it on real data.

Synthetic data can also be used to test and validate non-routine event detection and outlier rejection algorithms. M&V 2.0 algorithms can have challenges in detecting non-routine events—such as temporary changes in operating hours, special events, or construction—so as to exclude them from the algorithms' training period. Adding non-routine events to synthetic data and having ground truth information for when these events take place can help test and validate M&V 2.0 and non-routine event detection algorithms.

Synthetic meter data can be used to test and validate simulation model calibration as well. Again, the ground truth of an actual building model is unknown in real-world calibrations. By creating synthetic data and calibrating a model using that data, the input parameters that were calibrated (e.g., hours of operation, U-value of the walls, HVAC equipment efficiency) can be compared to the synthetic metadata to determine the accuracy of the calibration model or algorithm. This is useful in determining whether the calibrated model is successful in calculating accurate outputs using correct calibrated parameters, or whether the outputs are accurate but calculated based on incorrect input parameters.

5.2 Future work

Future areas to explore in synthetic data generation include methods to generate higher resolution output that better approximate real data. Current synthetic data is generated at a 15-minute interval; this resolution can be increased to represent higher-frequency events in real data, such as equipment cycling. Other use cases may need submetering data at the end-use level, which can be generated in the future.

Another area to improve in this study is to collect and analyze large-scale occupant activities and energy end-uses in real buildings across more building types and climate zones to provide better models of occupant-driven loads used in the algorithms to generate synthetic meter data dynamics. Stochastic occupant behavior is one major reason behind the gap between simulated and real high-temporal-resolution energy consumption data.

Validation of synthetic meter data against real building meter data is also a future activity. Traditionally, metrics such as coefficient of variation of the root mean square error (CVRMSE) and the mean bias error (MBE) have been used to evaluate how similar two energy consumption profiles are. Those metrics are often used in model calibration and

M&V applications (Guideline 2014; Ruiz and Bandera 2017; Deme Belafi et al. 2019). However, those metrics are insufficient to evaluate how well the synthetic meter data agree with real meter data at a finer granularity (e.g., hourly or sub-hourly interval). This is because we are interested in whether the synthetic and real meter data have similar dynamics and variations, rather than whether they have exactly same values at same times. Therefore, other existing or new metrics and methods that can better compare the dynamics and variations of the synthetic and real load profiles will be explored and tested in future validation activities. A follow-up paper will present the validation methodology and results.

6 Conclusion

Synthetic meter data and associated synthetic metadata can be valuable data assets to support testing and validation of energy algorithms that target improvements in building energy efficiency. In this study, we proposed a framework to generate and represent the synthetic smart meter data. For meter data generation, we proposed assumptions on stochastic and dynamic occupant behavior models through multiple data sources; implemented those assumptions through OpenStudio measures; and then ran batch simulations to demonstrate the workflow. For data representation and query, we summarized existing smart meter data and metadata schemas and reviewed synthetic data storage and retrieval method. The proposed framework and the open-source OpenStudio Gems and measures will be made available at GitHub to provide an easy way for users to generate a large amount of synthetic smart meter data, supporting their building energy research and projects to reduce energy use and greenhouse gases emissions in buildings.

Acknowledgements

This research was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Building Technologies of the United States Department of Energy, under Contract No. DE-AC02-05CH11231.

References

- Andersen RK (2012). The influence of occupants' behaviour on energy consumption investigated in 290 identical dwellings and in 35 apartments. In: Proceedings of the 10th International Conference on Healthy Buildings.
- Bartz K (2019). Record wildfires push 2018 disaster costs to \$91 billion. Center for Climate and Energy Solutions. Available at <https://www.c2es.org/2019/02/record-wildfires-push-2018-disaster-costs-to-91-billion/>. Accessed 1 Nov 2019.

- BEDES (n.d.). Building Energy Data Exchange Specification (BEDES). Available at <https://bedes.lbl.gov/>. Accessed 23 Sept 2019.
- BrickSchema (n.d.). Available at <https://brickschema.org/>. Accessed 29 Sept 2019.
- BuildingSync (n.d.). Available at <https://buildingsync.net/>. Accessed 29 Sept 2019.
- Chao CYH, Hu JS (2004). Development of a dual-mode demand control ventilation strategy for indoor air quality control and energy saving. *Building and Environment*, 39: 385–397.
- Chen Y, Hong T, Luo X (2018). An agent-based stochastic Occupancy Simulator. *Building Simulation*, 11: 37–49.
- D'Oca S, Hong T, Langevin J (2018). The human dimensions of energy use in buildings: A review. *Renewable and Sustainable Energy Reviews*, 81: 731–742.
- Dahmen J, Cook D (2019). SynSys: A synthetic data generation system for healthcare applications. *Sensors*, 19: 1181.
- de Bakker C, Aries M, Kort H, Rosemann A (2017). Occupancy-based lighting control in open-plan office spaces: A state-of-the-art review. *Building and Environment*, 112: 308–321.
- de Dear RJ (1998). A global database of thermal comfort field experiments. *ASHRAE Transactions*, 104(1b): 1141–2252.
- Deme Belafi Z, Hong T, Reith A (2019). A library of building occupant behaviour models represented in a standardised schema. *Energy Efficiency*, 12: 637–651.
- Deru M, Field K, Studer D, Benne K, Griffith B, et al. (2011). U.S. Department of Energy Commercial Reference Building Models of the National Building Stock. Office of Scientific and Technical Information (OSTI).
- Dickinson R (2016). Morphing Weather Files: An Overview of the Weathershift Tool. Available at <https://www.iesve.com/website/var/assets/support/weather-files/weathershift/weathershift-white-paper.pdf>.
- Dunn G, Knight I (2005). Small power equipment loads in UK office environments. *Energy and Buildings*, 37: 87–91.
- Fernandes LL, Lee ES, DiBartolomeo DL, McNeil A (2014). Monitored lighting energy savings from dimmable lighting controls in The New York Times Headquarters Building. *Energy and Buildings*, 68: 498–514.
- Field K, Deru M, Studer D (2010). Using DoE Commercial Reference Buildings for Simulation Studies. In: Proceedings of SimBuild. Available at Retrieved from <http://www.nrel.gov/docs/fy10osti/48588.pdf>.
- Földváry Ličina V, Cheung T, Zhang H, de Dear R, Parkinson, et al. (2018). Development of the ASHRAE Global Thermal Comfort Database II. *Building and Environment*, 142: 502–512.
- Guglielmetti R, Macumber D, Long N (n.d.). OpenStudio: An Open Source Integrated Analysis Platform.
- Guideline A (2014). Guideline 14-2014. Measurement of Energy, Demand, and Water Savings.
- Gunay HB, O'Brien W, Beausoleil-Morrison I, Gilani S (2016). Modeling plug-in equipment load patterns in private office spaces. *Energy and Buildings*, 121: 234–249.
- Guo X, Tiller DK, Henze GP, Waters CE (2010). The performance of occupancy-based lighting control systems: A review. *Lighting Research & Technology*, 42: 415–431.
- Haves P, Parker A, Jegi S, Garg V, Ravache B (2017). Development of automated procedures to generate reference building models for ASHRAE Standard 90.1 and India's building energy code and implementation in OpenStudio.
- Hong T, Chang W-K, Lin H-W (2013). A fresh look at weather impact on peak electricity demand and energy use of buildings using 30-year actual weather data. *Applied Energy*, 111: 333–350.
- Hong T, Lin H-W (2013). Occupant behavior: impact on energy use of private offices. LBNL Report, LBNL-6128E. Lawrence Berkeley National Laboratory.
- Hong T, Yan D, D'Oca S, Chen CF (2017). Ten questions concerning occupant behavior in buildings: The big picture. *Building and Environment*, 114: 518–530.
- Kim Y-S, Heidarinejad M, Dahlhausen M, Srebric J (2017). Building energy model calibration with schedules derived from electricity use data. *Applied Energy*, 190: 997–1007.
- Kim Y-S, Srebric J (2017). Impact of occupancy rates on the building electricity consumption in commercial buildings. *Energy and Buildings*, 138: 591–600.
- Knight W (2016). Self-driving cars can learn a lot by playing grand theft auto. *MIT Technology Review*, Available at <https://www.technologyreview.com/s/602317/self-driving-cars-can-learn-a-lot-by-playing-grand-theft-auto/>. Accessed 23 Sept 2019.
- Li H, Macumber D (2019). Open studio-Occupant-Variability-Gem: Pre-release of OpenStudio-Occupant-Variability-Gem. Available at <https://doi.org/10.5281/zenodo.3458596>.
- Luo X, Lam KP, Chen Y, Hong T (2017). Performance evaluation of an agent-based occupancy simulation model. *Building and Environment*, 115: 42–53.
- Mahdavi A, Tahmasebi F, Kayalar M (2016). Prediction of plug loads in office buildings: Simplified and probabilistic methods. *Energy and Buildings*, 129: 322–329.
- Marr B (2018). Does synthetic data hold the secret to artificial intelligence? *Forbes*, Available at <https://bit.ly/2ne9bSE>. Accessed 23 Sept 2019.
- Martani C, Lee D, Robinson P, Britter R, Ratti C (2012). ENERNET: Studying the dynamic relationship between building occupancy and energy consumption. *Energy and Buildings*, 47: 584–591.
- Masoso OT, Grobler LJ (2010). The dark side of occupants' behaviour on building energy use. *Energy and Buildings*, 42: 173–177.
- Mathew P, Wallace N, Issler P, Ravache B, Sun K, Coleman P, Zhu C (2018). Do energy costs really affect commercial mortgage default risk? New results and implications for energy efficiency investments.
- Motegi N, Piette MA, Watson DS, Kiliccote S, Xu P (2007). Introduction to commercial building control strategies and techniques for demand response—Appendices. Office of Scientific and Technical Information (OSTI).
- Nikolaev EI, Dvoryaninov PV, Lensky YY, Drozdovsky NS (2018). Using virtual data for training deep model for hand gesture recognition. *Journal of Physics: Conference Series*, 1015: 042045.
- Project Haystack (n.d.). Available at <https://project-haystack.org/>. Accessed 29 Sept 2019.
- Ranson M, Tarquinio L, Lew A (2016). Modeling the impact of climate change on extreme weather losses. Available at <https://doi.org/10.22004/AG.ECON.280932>.

- Roth A, Goldwasser D, Parker A (2016). There's a measure for that!. *Energy and Buildings*, 117: 321–331.
- Ruiz G, Bandera C (2017). Validation of calibrated energy models: common errors. *Energies*, 10: 1587.
- Sarkar T (2018). Synthetic data generation—A must-have skill for new data scientists. Available at <https://towardsdatascience.com/synthetic-data-generation-a-must-have-skill-for-new-data-scientists-915896c0c1ae>. Accessed 23 Sept 2019.
- Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R (2017). Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition.
- Thomas D, Butry D, Gilbert S, Webb D, Fung J (2016). The costs and losses of wildfires: a literature survey. National Institute of Standards and Technology. Available at <https://doi.org/10.6028/NIST.SP.1215>.
- Tian Y, Li X, Wang K, Wang F (2018). Training and testing object detectors with virtual images. *CAA Journal of Automatica Sinica*, 5: 539–546.
- Toole J (2019). Synthetic data: A bridge over the data moat—Jameson Toole—Medium. Available at <https://heartbeat.fritz.ai/synthetic-data-a-bridge-over-the-data-moat-29f392a52f27>. Accessed 23 Sept 2019.
- Wang Z, de Dear R, Luo M, Lin B, He Y, Ghahramani A, Zhu Y (2018). Individual difference in thermal comfort: A literature review. *Building and Environment*, 138: 181–193.
- Wang N, Vlachokostas A, Borkum M, Bergmann H, Zaleski S (2019). Unique Building Identifier: A natural key for building data matching and its energy applications. *Energy and Buildings*, 184: 230–241.
- Wang Z, Hong T (2020). Learning occupants' indoor comfort temperature through a Bayesian inference approach for office buildings in United States. *Renewable and Sustainable Energy Reviews*, 119: 109593.
- Wehner MF, Arnold JR, Knutson T, Kunkel KE, LeGrande AN (2017). Ch. 8: Droughts, floods, and wildfires. In: Climate Science Special Report: Fourth National Climate Assessment, Volume I. U.S. Global Change Research Program. Available at <https://doi.org/10.7930/J0CJ8BNN>.
- Wilcox S, Marion W (2008). Users Manual for TMY3 Data Sets (Revised). Office of Scientific and Technical Information (OSTI). Available at <https://doi.org/10.2172/928611>.
- Yan D, Hong T, Dong B, Mahdavi A, D'Oca S, Gaetani I, Feng X (2017). IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings. *Energy and Buildings*, 156: 258–270.
- Yoshino H, Hong T, Nord N (2017). IEA EBC annex 53: Total energy use in buildings—Analysis and evaluation methods. *Energy and Buildings*, 152: 124–136.