Titre: Title:	Development of a stochastic virtual smart meter data set for a residential building stock - methodology and sample data
Auteurs: Authors:	Adam Neale, Michaël Kummert, & Michel Bernier
Date:	2020
Type:	Article de revue / Article
Référence: Citation:	Neale, A., Kummert, M., & Bernier, M. (2020). Development of a stochastic virtual smart meter data set for a residential building stock - methodology and sample data. Journal of Building Performance Simulation, 13(5), 583-605. https://doi.org/10.1080/19401493.2020.1800096

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/5510/
Version:	Version finale avant publication / Accepted version Révisé par les pairs / Refereed
Conditions d'utilisation: Terms of Use:	Tous droits réservés / All rights reserved

Titre de la revue: Journal Title:	Journal of Building Performance Simulation (vol. 13, no. 5)
Maison d'édition: Publisher:	Taylor & Francis
URL officiel: Official URL:	https://doi.org/10.1080/19401493.2020.1800096
Mention légale: Legal notice:	This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of Building Performance Simulation (vol. 13, no. 5) in 2020, available online: https://doi.org/10.1080/19401493.2020.1800096

Development of a stochastic virtual smart meter data set for a residential building stock - methodology and sample data

Adam Neale*, Michaël Kummert and Michel Bernier

Département de génie mécanique, Polytechnique Montréal, Montréal, Canada

Adam Neale, Polytechnique Montréal, adam.neale@polymtl.ca.

Accepted manuscript

Journal of Building Performance Simulation 13 (5), 583-605

Development of a stochastic virtual smart meter data set for a residential building stock - methodology and sample data set

Existing electricity smart meter data sets lack sufficient details on building parameters to evaluate the impact that home characteristics can have on electricity consumption. An extensive, open-source virtual smart meter (VSM) data set with corresponding building characteristics is provided. The methodology used to develop the VSM data is presented in detail. The data set consists of a variety of homes representative of a subset of the Canadian single-family home building stock. The building characteristics cover a wide range of values that are based on probability distributions developed using a segmentation and characterization process. The resulting framework and VSM data set can be used by researchers to develop classification models, verify load disaggregation algorithms, and for a variety of other purposes.

Keywords: smart meter data; building energy simulation; residential;

Subject classification codes: include these here if the journal requires them

Introduction

Estimating the energy use of multiple buildings, such as at the city- or regional-scale, is commonly referred to as building stock energy modeling (BSEM) or urban building energy modeling (UBEM). BSEM can be accomplished with a number of well-documented techniques (Swan and Ugursal 2009). One such technique makes use of building archetypes to simplify the modeling approach through a process of segmentation and characterisation of the building stock characteristics (Reinhart and Cerezo Davila 2016).

Building archetypes have been developed for numerous cities and countries across the globe. One of the key issues facing archetype development is the lack of reliable and accurate information on the buildings (Booth, Choudhary, and Spiegelhalter 2012). The accuracy of the building stock model depends on the quantity and quality of available data. In addition, data sources vary widely at municipal, regional and federal levels, which makes it difficult to apply a single methodology across all building stocks.

This work is part of a broader study to provide a new methodology for using electricity smart meter data as a data source for building stock characterisation and segmentation. A key component of this research consists in the development of classification models using machine learning on smart meter data sets. With a sufficiently large smart meter data set with known building characteristics, classification models could be developed and used to determine the characteristics from smart meter data for a variety of regions. However, this approach depends on two factors: 1) sufficient smart meter data availability, and 2) building characteristics that are associated with each set of smart meter data. To discuss the possibility of classification modeling, smart meter data availability must first be addressed.

Smart meter data

As of 2017, over 770 million electricity smart meters have been installed globally (IEA 2019). This amount has been steadily increasing in recent years, in particular due to significant interest in the Asia-Pacific region. There are 79 million meters in the United States of America alone (IEA 2019). The global market in terms of installed units increased by over 12% from 2016 to 2017 as countries seek to convert their building stock to new metering technologies (IEA 2017).

In the province of Québec, Canada, there are over 3.7 million electricity smart meters (Hydro-Québec 2016) currently installed, primarily for the 3.6 million household residential market (2016 data, Natural Resources Canada 2019a). These meters collect data at 15-minute intervals and are mostly used for billing purposes (Hydro-Québec 2012).

Smart meters are very prevalent and could present an interesting opportunity to extract information about residential energy consumption. While numerous applications are possible for analysing smart meter data, the authors focus primarily on developing classification models of residential smart meter data based on known building characteristics.

Supervised machine learning classification of smart meter data

While the purpose of this paper is not to discuss classification model development, a brief mention is made here for context. Supervised machine learning classification can be performed on electricity consumption data to estimate building parameters, so long as the classification algorithm has sufficient data to train the model. Unsupervised machine learning is typically unsuited for this process, as there is limited information on a building that can be extracted from electricity consumption data without training the model.

Some classification studies have been performed on real smart meter data (see e.g. Beckel et al. 2014; Carroll et al. 2018; Neale et al. 2019). To classify smart meter data requires:

1) Electricity consumption at a given sampling rate (classification predictor);
2) Known building parameters (classification response).
3) An appropriate classification algorithm (e.g. discriminant analysis).

Since the objective of the authors is to use classification on smart meter data, it is important to discuss available smart meter data sets.

Electricity smart meter data available for classification studies

There are very few open-source residential electricity meter data sets with known building characteristics. Those that do exist have very little information on the buildings. Utilities are often reluctant to share the meter data due to privacy reasons. Of the electricity data sets that were found, the works were divided into two common themes: 1) smart meter data for occupant behaviour analysis, and 2) high-resolution submetered electricity for load disaggregation studies. Data sets were researched based on their location, number of homes in the sample, sampling frequency, trial duration and relevant building characteristics, which are presented in Table 1.

Table 1. Summary of open-source residential smart meter data sets with relevant building information

Data set	Loc.	# homes	Sampling period	Trial duration	Building information	Ref.
Smart meter data sets						
CER electricity customer behaviour trial	Ireland	4232	30 minutes	1.5 years	Occupant social data, appliance use, some building geometry information.	CER (2012)
PNNL GridWise Demonstration Project	USA	112	15 minutes	~1 year	Occupant surveys	Hammerstrom, Ambrosio et al. (2007), Hammerstrom, Brous et al. (2007)
Load disaggregation data sets						
UMass Smart* Home Dataset (2017 release)	USA	7	1 minute	Varies (1-2 years)	Weather data, very detailed submetered electricity consumption.	Barker et al. (2012)
UMass Smart* Microgrid Dataset	USA	443	1 minute	24 hours	Electricity consumption only.	Barker et al. (2012)
REFIT smart home dataset	UK	20	Mixed	2 years	Building occupant survey data, complete building description, high-resolution appliance electricity use.	Murray et al (2017).
Almanac of Minutely Power dataset 2	Canada	1	1 minute	2+ years	Building geometry	Makonin et al. (2016)
Dutch Residential Energy Dataset	The Netherlands	1	1 second	6 months	Building geometry, occupancy, appliances, indoor temperature	(Nambi et al. 2015)
ECO dataset	Switzerland	6	1 second	8 months	Occupancy data	Kleiminger et al. (2015)
Carleton high-resolution electricity data set – Study 1	Canada	12	1 minute	~14 months	Type, vintage, building surface area, number of occupants	Saldanha & Beausoleil-Morrison (2012)
Carleton high-resolution electricity data set – Study 2	Canada	23[1]	1 minute	~1 year	Type, vintage, building surface area, number of occupants	Johnson & Beausoleil-Morrison (2017)

[1] Includes the 12 homes from the study by Saldanha & Beausoleil-Morrison (2012).

The smart meter data sets are characterized by a relatively high number of homes and a monitoring period exceeding 1 year. They focused on occupancy behaviour and contained little or no details on the characteristics of the buildings. The Irish Commission for Energy Regulation (CER) electricity customer behaviour trial

(CER 2012) is the best example of open-source smart meter data with over 4000 single-family homes. Relevant building characteristics include the type of building (detached, semi-detached), building floor area and number of occupants. The CER data set does not include any heating or cooling electricity use as these Irish homes tend to have no cooling system and have non-electric heating (Beckel et al. 2014).

Load disaggregation data sets are included in this review since they typically contain more detailed information on the building's characteristics and the high-resolution electricity consumption data can be aggregated. Due to the sheer quantity of data and measurement points, load disaggregation studies typically limit the scope of the monitoring to a smaller number of homes and/or shorter monitoring period. There is therefore little diversity in house types for these types of studies. The UMass Smart Microgrid Dataset has the highest sample of buildings with 443 homes, however it only contains the electricity consumption data and no building information (Barker et al. 2012).

The available data sets are therefore inadequate for developing classification models based on the electricity consumption at sampling rates representative of electricity smart meter data. There is either too limited information on the building to classify data, too few homes to represent the diversity of the building stock, or too little data to represent a full year of electricity consumption.

Objectives

The principal objective of the authors is to estimate building characteristics from anonymous electricity smart meter data, an approach previously described in Neale et al. (2018). While this would normally be accomplished by training classification models using a smart meter data set with known building parameters, no such data sets are extensive enough to perform that task at this stage. The main objective therefore is to develop a virtual smart meter (VSM) data set with known building characteristics using batched building energy simulations. The VSM data set will be used to develop a clear link between building characteristics and electricity consumption at actual smart meter data resolution.

The single-family home market is targeted as it is a significant portion of the residential building stock in Canada, representing 16.5% of secondary energy use (2017 data, NRCan 2020). Expansion to other markets is possible in future work. Given the need to train classification models for many different building characteristics, the building characteristics must be generated in a way to be as representative as possible of the building stock, to minimize combinations of parameters that are not likely to appear in the chosen market.

The developed framework must therefore have the capability to:

- Generate single-family homes using building energy simulation;
- Use building characteristics and conditions typical of the chosen market using probability distributions, wherever possible;
- Determine the electricity consumption of the home, effectively producing a “virtual smart meter data profile”;
- Link the building characteristics to the virtual smart meter data profile for each generated building;
- Generate a large quantity of homes (e.g. 100 000+ homes) to best cover the range of possible building parameters for classification model development, i.e. via batch simulation.

Methodology

The methodology used to develop the framework and produce the VSM data set for an example building stock is presented in a way that can facilitate applying it to other building stocks. A sample set of VSM data with the corresponding building characteristics is provided as supplementary material with the paper. This open-source data set can be used to test classification algorithms and study the impact of various building parameters and occupancy profiles on electricity use at typical smart meter sampling periods.

The framework concept is divided into two main components illustrated in Figure 1: 1) *Generator*, which is the component that produces the VSM data profiles, and 2) *Classifier*, which is the classification model module. While the *Generator* component is the focus of this paper, it is relevant to present the *Classifier* module in part to justify certain choices made in the development of the former.

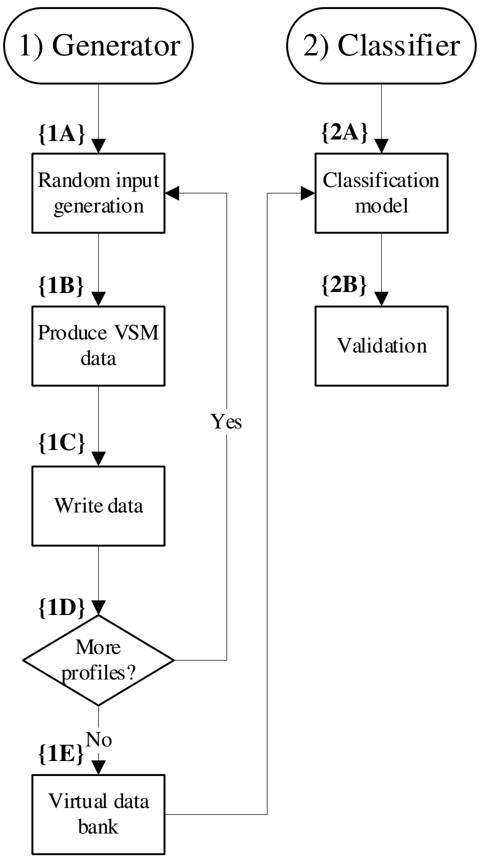


*Figure 1. Virtual smart meter data generation (*Generator*) and classification (*Classifier*) processes*

The *Generator* component consists of five parts, described as {1A} to {1E} in Figure 1. First, the building parameters are generated randomly {1A} to ensure a unique building is generated upon each iteration of the model. This is accomplished by generating building characteristics according to predetermined probability distributions. A building simulation is then performed {1B} and the data are output in a specific format corresponding to typical smart meter data {1C}. The module then determines whether additional profiles are required {1D} and the process would continue. If the data generation is complete, the VSM profiles are compiled into a single data bank for future use {1E}.

The *Classifier* component then reads the data bank and develops a classification model based on the smart meter data and known building parameters {2A}. The result would then be validated using real smart meter data for the targeted building stock, i.e. the building characteristics for the real smart meter data would be compared with predicted values from the classification process {2B}.

Virtual smart meter data generation

Virtual smart meter (VSM) data is defined here as electricity consumption data generated using a physics-based building energy model at 15-minute intervals, though other sampling rates are possible. The sampling rate selected for the virtual data should match the one used by the local electricity distributor.

VSM data is intended to replicate the electricity consumption for single-family homes of various compositions. While the methodology behind the model can be applied to any region, the characteristics selected are representative of homes found in the province of Québec, Canada.

Development of a virtual smart meter framework

As illustrated in Figure 1, the purpose of the developed framework is to stochastically generate a set of building parameters and then execute a building energy simulation using those parameters. The framework produces a VSM data profile that consists of electricity consumption data (in kWh) at 15-minute intervals, which is 35 040 data per home for a year. Each data profile is paired with the building parameters that are used to generate it. This process is then repeated the desired number of times to create a large data set for classification. As an example, the authors typically generate 200 000 homes with corresponding inputs and electricity consumption, which is well over 7 billion data.

The single-family home building stock for the province of Québec, Canada, is characterized by a variety of detached, semi-detached and row houses that typically have a basement. The homes are usually either one- or two-storeys and have a variety of thermal envelope performance levels and occupancy characteristics. Housing density (and therefore housing types) varies across the province, as does the climate. The framework therefore must include the following features:

1) Generating physical characteristics, such as size, shape, number of floors, for a home in the province of Québec;
2) Using a weather data file that represents the climate for each region of the province of Québec;
3) Occupying the virtual home with realistic occupants and internal loads and simulating their demands;
4) Producing the annual electricity use for the heating, cooling, lighting, appliances and domestic hot water loads of the home;
5) Repeating the process a large number of times with statistically representative inputs each time.

In order to generate a set of housing electricity consumption that is realistic, a framework had to be developed that could generate parameters that are representative for the chosen building stock and that correctly impact the house's energy use.

VSM framework details

The proposed framework consists of two main components: 1) a manager, which generates the building characteristics and starts a building energy simulation automatically, and 2) the building energy model, which generates the VSM data set given the set of selected parameters. The manager is an essential part of the process of automating the generation of building parameters and batch building simulations required to produce a significant set of VSM data. The logic behind the manager-building model interaction is illustrated in Figure 2.

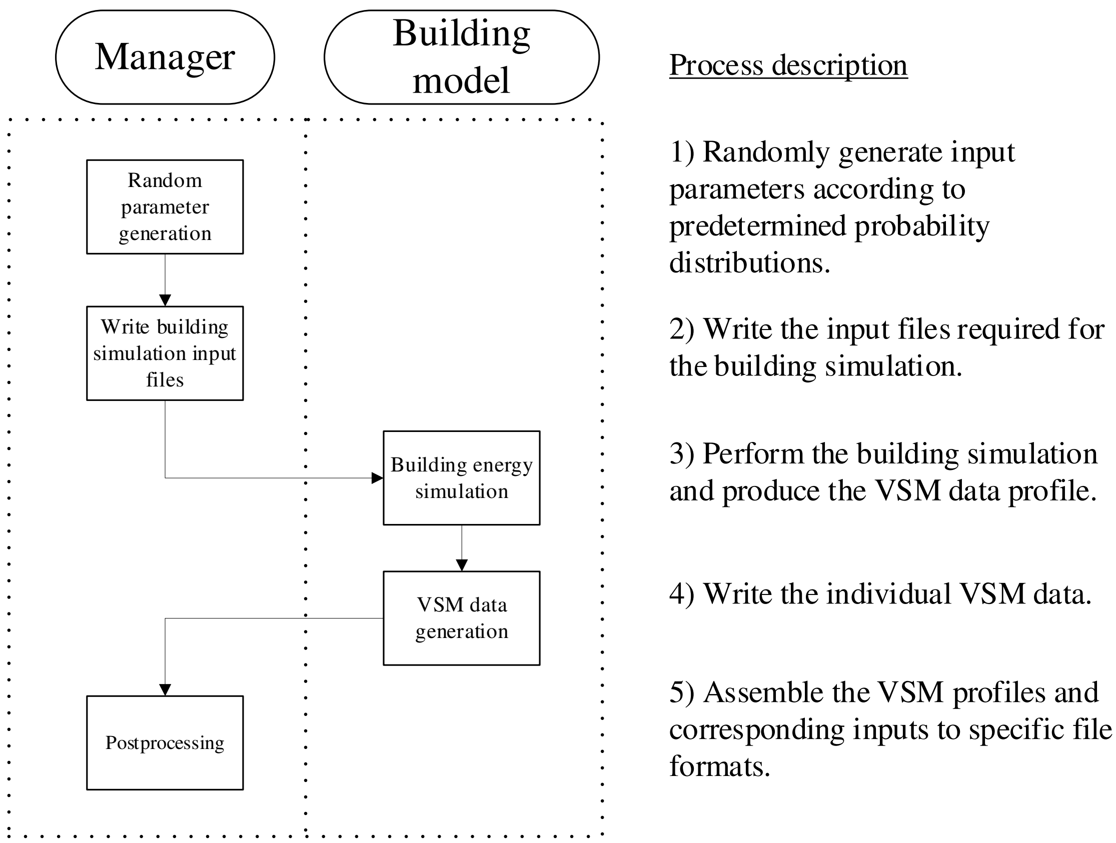


Figure 2. Proposed framework including a manager and building energy model

The manager is implemented in the Matlab environment (Mathworks 2018). The building model used is the Type 56 implemented within the TRNSYS environment (Klein et al. 2017). This model takes into account thermal mass to perform a transient simulation. Other dynamic simulation environments and building simulation programs could work equally well for the task. The building modeling approach for determining the building geometry, heating and cooling systems, lighting, appliance, domestic hot water and air infiltration loads are discussed in the following sections of the paper.

Building geometry

The building is represented as a single-zone rectangular prism. By default, the building is oriented with the south-facing wall as the "front" of the home, i.e. the façade that is facing the street. If the house type is semi-detached or a row house, one or both of the east- and west-facing walls are considered adjacent to another dwelling and not exposed to outdoor conditions. Houses can be one- or two-storeys. The heated surface area of a home is considered the sum of the floor areas including a basement. The building rotation is specified in order to determine the solar gains for each surface. The aspect ratio of the building's depth to width determines the area of each external surface. The height of each storey is considered constant (2.4 m).

External walls and roof surfaces are insulated with specified thermal resistance values. Windows are evenly distributed on all external envelope surfaces based on a window-to-wall ratio. Walls that are shared between more than one dwelling do not contain any windows, such as in semi-detached or row houses.

The building model includes a basement that is primarily below ground level, where the foundation and slab are insulated with a specified thermal resistance value. The ground temperature is modelled with a sinusoidal external boundary condition described in Equation (1).

$$T_{ground} = C_0 + C_1 cos(C_2 t + C_3) \quad (1)$$

where T_{ground} is the ground temperature on the outer surface of the building envelope (°C), C_i are constants, and t is time (h). Constants are specific to the region where the house is located and whether the external surface is a foundation wall or beneath a slab. Coefficients in equation 1 were pre-calculated for each region and determined based on a 3-D finite-difference type simulation to predict heat transfer in basement walls and slabs.

Heating systems

The heating system for a home can consist of one of three possible system configurations (note: symbols for equations appear at the end of this section):

(1) Electric element heating, i.e. electric baseboards or an electric furnace. The heating load for the home is calculated and a constant coefficient of performance (COP) is applied to calculate the electric heating requirement. $COP_{electric} = 1.0$, $PF = 1.0$.

$$E_{Heat,elec} = \frac{PF \times Q_{Heat,load} \times \Delta t}{COP_{electric}} \quad (2)$$

(2) Non-electric heating, i.e. a natural gas or heating oil furnace. The heating load for the home is calculated and a parasitic load is applied for electric subsystems and applied in the form of a power fraction (PF). $COP_{electric}$ = 1.0, PF = 0.05.

$$E_{Heat,elec} = \frac{PF \times Q_{Heat,load} \times 0.25}{COP_{electric}} \quad (3)$$

(3) Air-source heat pump (ASHP) with an auxiliary (AUX) heating system. In this case, the heating system transitions between the ASHP and AUX systems based

on the outdoor temperature. The usage fraction for the ASHP (F_{ASHP}) and AUX (F_{AUX}) are expressed using the following heuristic relationships:

For $T_{outdoor}$ > -5 °C

$$F_{AUX} = 0$$

$$F_{ASHP} = 1.0$$

For -12 °C ≤ $T_{outdoor}$ ≤ -5 °C

$$F_{AUX} = 0.1429 \times T_{Outdoor} + 1.7143$$

$$F_{ASHP} = 1 - F_{AUX}$$

For $T_{outdoor}$ < -12 °C

$$F_{AUX} = 1.0$$

$$F_{ASHP} = 0$$

The COP for the ASHP is determined as a function of outdoor air-temperature, as depicted in Equation (4).

$$COP_{ASHP} = 0.0585 \times T_{Outdoor} + 3.115 \tag{4}$$

The overall heating system electricity usage for a heat pump system with auxiliary heating for a given time step is expressed using Equation (5).

$$E_{Heat,elec} = \frac{PF \times F_{AUX} \times Q_{Heat,load} \times \Delta t}{COP_{AUX}} + \frac{F_{ASHP} \times Q_{Heat,load} \times \Delta t}{COP_{ASHP}} \tag{5}$$

Description for symbols used in Equations (2) through (5):

- Δt is the time step in the simulation (0.25 h).
- COP_{AUX} is the COP of the auxiliary heating system.
- COP_{ASHP} is the COP of the air-source heat pump for the given time step.
- $E_{Heat,elec}$ is the amount of electricity required to heat the home for the given time step (kWh).
- F_{AUX} is the usage fraction for an auxiliary (AUX) heating system.
- F_{ASHP} is the usage fraction of an air-source heat pump (ASHP).
- PF is the power fraction for non-electric equipment.
- $Q_{Heat,load}$ is the heating load calculated by the building model (kW).
- $T_{outdoor}$ is the outdoor dry bulb temperature (°C).

Cooling systems

The electricity use for cooling a home is determined based on the presence of a cooling system. If no cooling system is present, the cooling electricity is zero. If an air-conditioner or a reversible heat pump exists, the cooling electricity use is modelled using Equation (6).

$$E_{Cool,elec} = \frac{Q_{Cool,load} \times \Delta t}{COP_{ASHP}} \tag{6}$$

where $E_{Cool,elec}$ is the amount of electricity required for cooling for the given time step (kWh), Δt is the duration of a time step (0.25 h), and COP_{ASHP} is the coefficient of performance calculated as a function of outdoor temperature, as expressed in Equation (4).

Lighting internal gains

Lighting is modelled as an internal heat gain corresponding to a heat source expressed in watts (W). Lighting heat gains to the surrounding environment are considered 57% radiative and 43% convective (ASHRAE 2013).

Equipment internal gains

Equipment internal gains are applied as an internal heat source expressed in watts (W). Heat gains to the surrounding environment are considered 30% radiative and 70% convective (ASHRAE 2013). Note that some equipment, such as swimming pool and spa pumps and heaters, are not typically installed within the home. While these devices are included in the total electricity consumption, they are not included as internal gains within the house.

Domestic hot water

Domestic hot water electricity use is determined using a hot water tank simulation using the TRNSYS software (Klein et al. 2017). The water heater model consists of a vertical cylindrical insulated tank that is equipped with master-slave heating elements, controlled with aquastats and with a volume equal to 266 L. Standby losses are determined by assuming a constant ambient air temperature and constant thermal resistance of the tank. Electricity use is calculated based on the activation of the heating elements due to temperature changes in the hot water tank subsequent to hot water draws.

Air infiltration

The Sherman-Grimsrud infiltration model is used to represent air infiltration, which determines the air changes per hour (ACH) as a function of the indoor and outdoor temperatures, the leakage area of the building envelope, the wind speed given the height of the building, and the pressure differences due to stack effect (Sherman and Grimsrud 1980).

Building parameters

A residential single-family building can be described using a variety of deterministic and probabilistic parameters. The number of virtual buildings required in the data set depends on the parameters being varied and the discretization of each for classification purposes. In other words, each characteristic is divided into discrete “bins” that are subsequently used for classification model development. Parameters are generally divided into 2 to 5 bins to ease the classification model development (Neale et al. 2019). In brief, if the goal is to correctly predict the category for a given building parameter, the likelihood of a correct prediction increases with fewer categories. The exactitude of the prediction is up to the classification modeler and the desired accuracy of the developed model.

A general description of the building parameters is provided in Table 2. The impact of the parameter on the building energy model is presented. The information available for each building stock will vary, though some examples of potential data sources are provided for each building parameter. Examples of categories for each characteristic are also provided in Table 2.

Table 2: Model inputs and potential data sources.

Property	Impact on model	Potential data sources	Categories/bins[1]
Location	Distribution of building types, climate determines heating/cooling degree days	National census data, national energy use databases	Number of locations depends on region studied. User choice. (7)*
Building type	Determines building geometry	National census data, building energy surveys, municipal tax evaluation data	Single-detached, semi-detached, row are the usual categories. Semi-detached and row can be combined. (2-3)
Shape	Determines building geometry	Building surveys, national studies, engineering knowledge of construction practices	Aspect ratios, e.g. 0.8, 1.0, 1.3. User choice. (3-5)
Rotation	Determines building geometry, solar gains	Engineering knowledge, map data	90° rotation increments. (4)
# Floors	Determines building geometry	Building surveys, national studies, engineering knowledge of construction practices	One or two floors. (2)
Wall construction	Determines building envelope thermal performance	Building surveys, national studies, engineering knowledge of construction practices	Wall thermal resistance levels. (4-5)
Roof construction	Determines building envelope thermal performance	Building surveys, national studies, engineering knowledge of construction practices	Roof thermal resistance levels. (4-6)
Foundation construction	Determines building envelope thermal performance	Building surveys, national studies, engineering knowledge of construction practices	Foundation thermal resistance levels. (4)
Infiltration	Impacts heating and cooling demand	Building surveys, national energy code levels, measurement campaigns	Rates of infiltration (3-5)
Window type	Determines building envelope thermal performance	Building surveys, national studies, engineering knowledge of construction practices	Multiple variations of each type are possible, but are categorized by number of glazings (Single, double, triple.). (3)

Property	Impact on model	Potential data sources	Categories/bins[1]
Window-to-wall ratio	Determines building envelope thermal performance and building geometry	Building surveys, national studies, engineering knowledge of construction practices	Surface area ratio window:wall, e.g. 0.1, 0.2, etc. (3)
Basement	Determines building geometry	Building surveys, national studies, engineering knowledge of construction practices	Yes/no/crawl space. (1-3)
Heating, ventilation and air conditioning	Determines heating, cooling and ventilation electricity use based on energy demand and fresh air needs	Building surveys, national studies, national energy use databases, equipment distributors.	Electric baseboards, central air, heat recovery ventilator, air conditioning. (2-3) each for cooling, heating, ventilation.
Setpoints	Determines heating and cooling demand	National studies, national energy use databases	Thermostat setpoints for heating and cooling, e.g. heating: 21°C, cooling: 25 °C. (1-3)
Appliances	Determines appliance electricity use	National studies, national energy use databases	Number and type of appliances (N/A)[2]
Lighting	Determines lighting electricity use	National studies, national energy use databases	Type of lighting, density. (N/A)[2]
Occupants	Impacts internal loads of the home	National census data, national studies	Number of occupants (5+)
Occupancy	Impacts internal loads of the home	Occupancy studies	Occupant schedule and activity schedule. (N/A)[2]
DHW consumption	Determines domestic hot water electricity use	National studies, research	Volume of hot water (N/A)[2]
Pool/spa installations	Determines electricity use due to a pool and/or spa installation	Equipment distributors, national studies, building surveys, building permits	Yes/no (2) each for pool and spa

[1] Values in parentheses represent the authors' recommended number of categories. Values with ( )* depend on the chosen building stock.
[2] As described in the present paper, internal loads are calculated stochastically using an independent tool and provided as an input to the building energy model.

A segmentation and characterization process of the Québec building stock was performed in order to fill out the categories described in Table 2. Data sources included Canadian census data (StatCan 2011; StatCan 2016), the Canadian Survey of Household Energy Use (NRCan 2011), and the Energuide Housing Database (NRCan 2018). Similar information may not be available for all regions, in which case engineering knowledge can be sufficient to define an appropriate distribution for each parameter. Based on the available information, parameters were divided into four categories:

1) Occupancy-driven internal loads;

2) Uniform probability distributions, which describe a set of parameter categories with equal probability;
3) Probability mass functions, which describe a set of unequal probabilities for a number of different categories for a given parameter.
4) Fixed parameters, which were input to the model as constant values for all building simulations.

Occupancy-driven internal loads

The number of occupants in a home can be established using a variety of data sources, which are discussed in subsequent sections of the paper. Of the building parameters described in Table 2 there are a number that are dependent on the occupancy of a home, i.e. the number of occupants that are at home and active in the house at any given time. Lighting, appliances and domestic hot water are primary examples of these loads.

To ensure a variety of stochastic occupancy behaviour that is directly tied to the internal loads of a home, the CREST thermal model (McKenna and Thomson 2016) is used to produce distinct internal load profiles for use in the *Generator* module. The CREST model is intended to produce daily profiles at 1-minute intervals, but was adapted to produce annual occupancy activity schedules with corresponding lighting, appliance and domestic hot water usage data at 15-minute intervals. The CREST model distinguishes between weekend and weekday behaviour but is limited to a maximum of 5 occupants. The presence of appliances (i.e. number of televisions and electronic devices), their usage (amount of time the device is operated based on occupancy), the number of occupants present at home, and their activity level are generated stochastically by the model based on real time-of-use probability tables. This ensures a wide variety of stochastic occupancy behaviour.

An example of a typical occupancy schedule with the corresponding domestic hot water usage is provided in Figure 3 for two full days. Occupancy (number of people in the house) and activity (number of people active in the house) are depicted for a 3-person household. Occupants are typically present during the night but not active, and absent during the day but active, though this varies from day to day. It should be noted that only occupants present in the house can influence the internal loads, and the activity level for an occupant not present is for descriptive purposes only. Domestic hot water draws occur when occupants are both present and active. Similar trends occur for lighting and appliance loads, though they are not depicted in Figure 3. Many appliances have constant or periodic electricity draws that are independent of occupancy, such as refrigerators. These are also represented in the CREST tool.

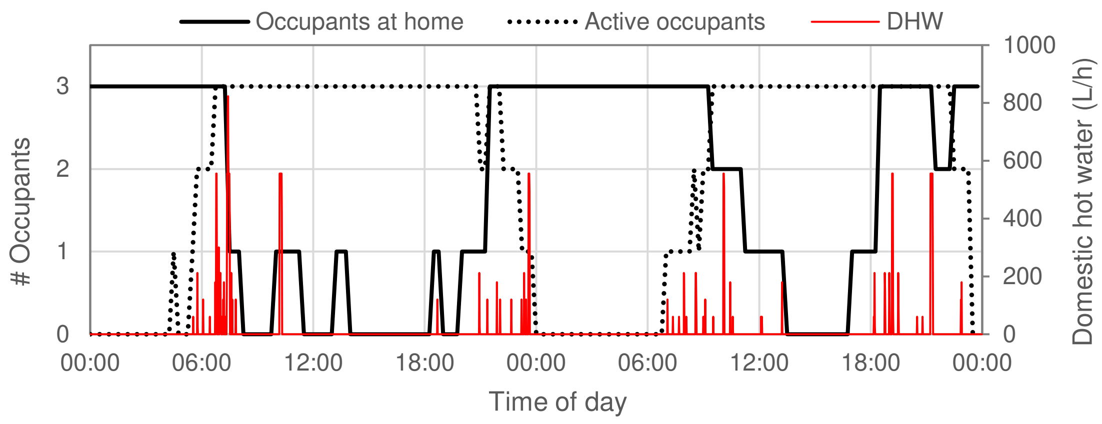


Figure 3. Example of an occupant activity schedule over a two-day period for a 3-occupant home.

To generate a variety of occupant behaviours, 15 stochastic occupancy schedules for 1 to 5 occupants are generated, for a total of 75 occupancy profiles. The stochastic internal load profiles are provided with the VSM data set for context. If time-of-use data is available for the studied building stock, then more specific internal load profiles could be generated by updating the CREST probability tables or by implementing the internal loads in a different manner. Since the schedule and number of occupants are treated as separate inputs, it is possible to generate houses with similar physical characteristics but different occupancy.

In addition to the loads above, swimming pools and portable electric spas represent a non-negligible electric energy usage in the province of Québec. Swimming pools utilize between 4300 kWh and 7500 kWh annually depending on whether they are above- or below-ground, and whether they are heated (Hydro-Québec 2019a). The fraction of houses with swimming pools varies depending on the type of dwelling. Pools are considered operational from June 1st to October 31st. Portable electric spas were found to use between 4500 kWh and 5500 kWh annually, depending on the frequency of use (Hydro-Québec 2019b). Spas are considered operational all year long.

Uniform probability distributions

A parameter described by a uniform probability distribution (UPD) has equal probability for all outcomes, as described in Equation (7).

$$P_{UPD}(A) = \frac{1}{k} \tag{7}$$

where $P_{UPD}(A)$ is the probability for a building parameter A with a uniform probability distribution and k is the number of categories for that parameter.

UPD are typically applied in cases where no prior probability data could be found for the studied building stock. For example, data was not available for the Québec residential building stock to characterize the frequency at which each rotation value occurs, therefore equal probability was assigned to rotation values equal to 0°, 90°, 180° and 270°. While it is preferable to apply a probability mass function where data allows, a UPD will nevertheless cover the range of possible values for a parameter allowing for correct classification. The disadvantage of using uniform probability distributions without correlation to other parameters is that occasionally there may be combinations that do not exist in the building stock or that are overrepresented, using computational resources for the classification process for no real benefit. A list of UPD parameters and their corresponding number of categories and probabilities are presented in Table 3.

Table 3. Building parameters with uniform probability distributions

Parameter	**Categories**	**Category description**	**Uniform probability P_{UPD}**	**Notes**
Rotation	4	Angle of rotation of the building with respect to true north. 0°, 90°, 180°, 270°	0.250	GIS[1] data could help characterize this parameter.
Shape	5	Ratio of width to length for a house. 0.8, 0.9, 1.0, 1.1, 1.2	0.200	Semi-detached houses can have asymmetrical configurations and it is therefore important to represent both aspect ratio and building rotation. GIS[1] data could help characterize this parameter.
Window-to-wall ratio	3	Ratio of window surface area to aboveground vertical building envelope area. 0.1, 0.15, 0.2	0.333	Applied to all vertical building envelope surfaces that are above ground level that are not directly adjacent to another home, i.e. shared surfaces for semi-detached or row houses.
Occupancy profile	15	Stochastically generated occupancy profiles. 1 to 15.	0.067	Individually generated occupancy profiles. 15 profiles were generated for each occupant category, i.e. 15 profiles for 1 occupant, 15 profiles for 2 occupants, etc.
Adjacent building surfaces for semi-detached homes	4	Determines which external building surface is adjacent to a neighbouring home. "None", "Both", "East" or "West".	0.250	Assuming that the front entrance of the building is facing south by default, the "east" and "west" terminology is adapted to describe which surface borders with a neighbour and therefore is not exposed to outdoor conditions.

[1] Geographic information systems

Probability mass functions

Probability mass functions (PMF) were established for a number of building parameters where statistical data was obtained. In many cases, sufficient data was available to establish dependence between one or more parameters, which is illustrated in Figure 4. The building location was selected first and subsequent parameters were chosen as a function of the location.

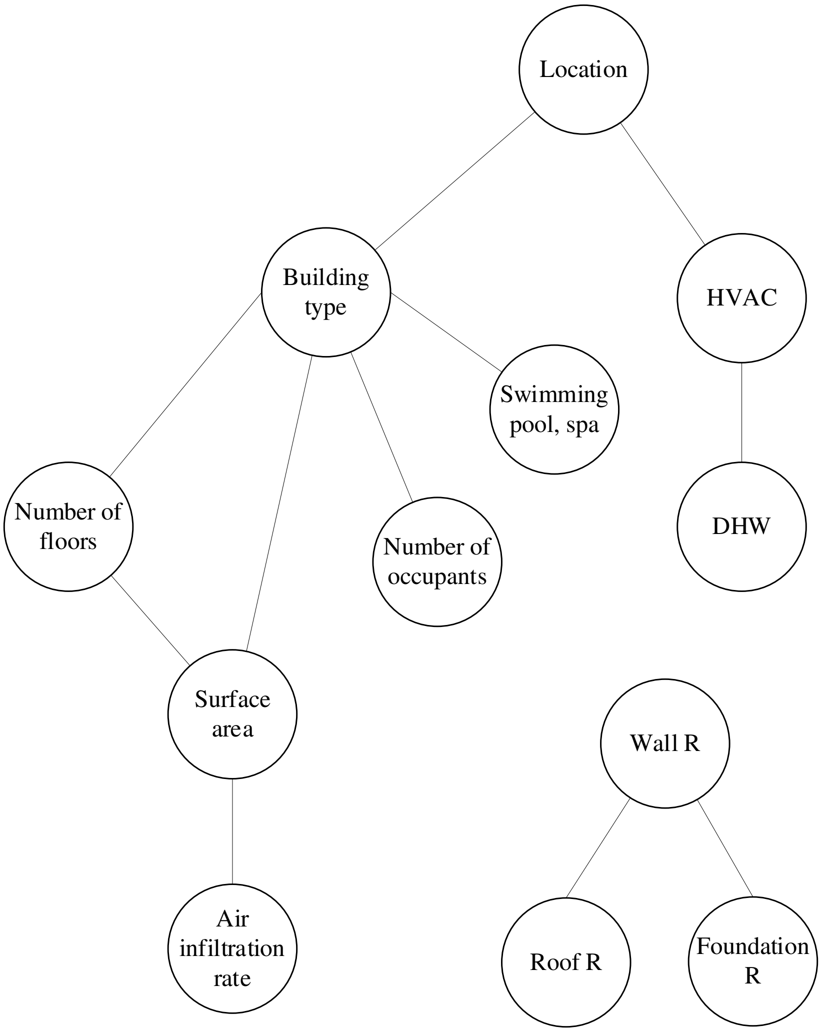


Figure 4. Building parameter dependency network. R: thermal resistance of the building envelope, DHW: domestic hot water, HVAC: heating, ventilation and air conditioning.

Bayes' theorem, described in Equation (8), is applied for each of the connections in the network in Figure 4, which allows for the determination of the conditional probability of a parameter given prior evidence.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{8}$$

where *A* and *B* are dependent parameters, $P(A|B)$ is the conditional probability of *A* given *B* occurring, $P(A)$ is the prior probability distribution of *A*, $P(B)$ is the prior probability distribution of *B*, and $P(B|A)$ is the prior conditional probability of *B* given *A* occurring. $P(B|A)$ is typically based on prior knowledge, i.e. based on data obtained in the literature. Prior probability distributions such as *P(A)* can be established based on evidence found in the literature and expressed using Equation (9).

$$P(A) = \frac{n_{Ai}}{n} \tag{9}$$

where n_{Ai} is the number of cases for class i of parameter A and n is the number of samples.

As an example of Bayes' Theorem applied to building parameters, the conditional probability distribution for P(Roof R|Wall R), which is read as *the probability of a building having a certain value of roof thermal resistance given a value of wall thermal resistance*, is illustrated in Table 4. The shaded values represent the maximum probability of roof thermal resistance for a given category of wall thermal resistance, i.e. the most likely possibility. Note that "*R*" denotes *thermal resistance*.

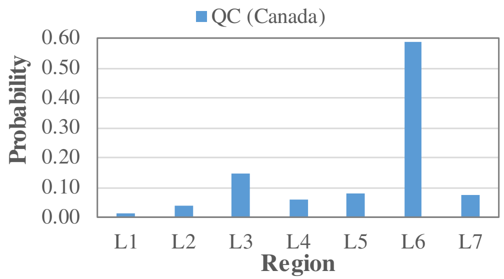
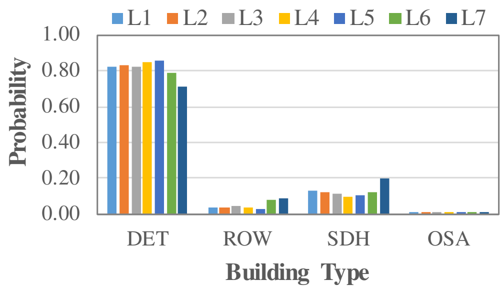
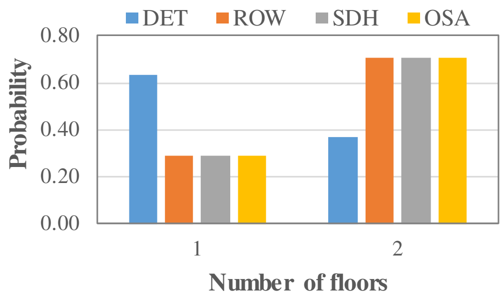
Table 4. Conditional probability distributions for roof thermal insulation based on wall insulation levels. Shaded values represent maximums.

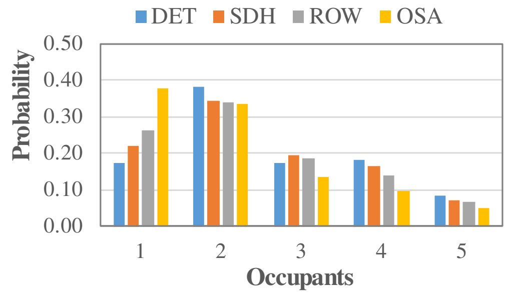
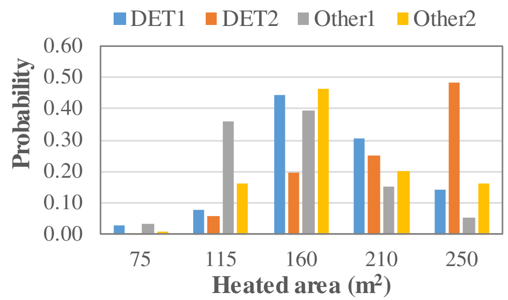
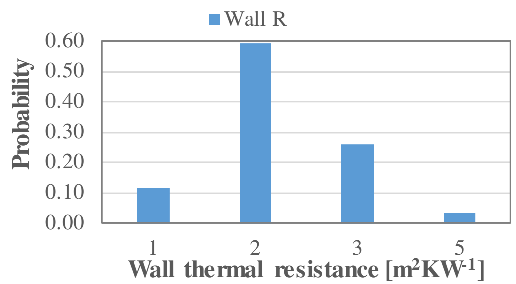
Roof R [$m^2 \cdot K \cdot W^{-1}$]	**Wall R** [$m^2 \cdot K \cdot W^{-1}$] **1**	**2**	**3**	**5**	**P(Roof R)**
1	0.146	0.041	0.008	0.008	0.043
2	0.319	0.250	0.069	0.068	0.204
3	0.207	0.214	0.109	0.096	0.181
4	0.138	0.194	0.193	0.134	0.186
5	0.095	0.142	0.295	0.190	0.179
8	0.095	0.159	0.326	0.504	0.208

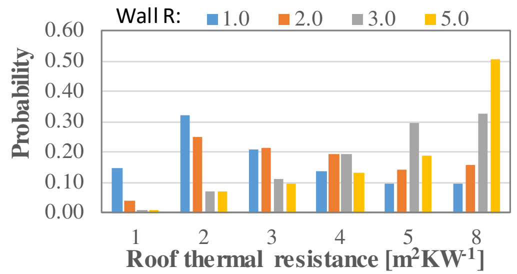
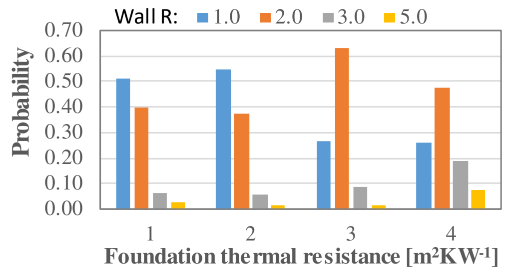
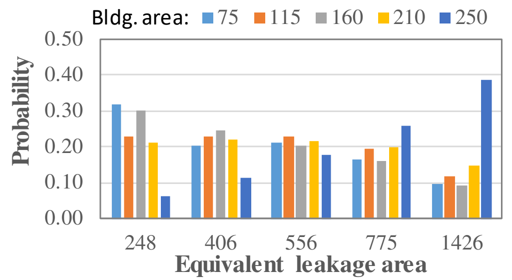
The conditional probability values in Table 4 illustrate the dependent nature of the roof insulation based on the level of the wall thermal resistance. For example, based on prior evidence, the probability of a building having a 2.0 $m^2 \cdot K \cdot W^{-1}$ roof insulation for a house with 2.0 $m^2 \cdot K \cdot W^{-1}$wall thermal insulation is 0.250. The higher the insulation in the walls, the more probable it is to have higher levels of roof insulation. A non-negligible portion of the building stock has unconventional configurations, which is important to represent correctly in the model. For example, approximately 9.5% of homes with very low wall thermal resistance (1.0 $m^2 \cdot K \cdot W^{-1}$) have very high roof thermal resistance (8.0 $m^2 \cdot K \cdot W^{-1}$). This is plausible due to the ease that attic insulation can be retrofitted to higher values of thermal insulation. The opposite is not true however, where homes with high values of wall thermal resistance and low values of roof thermal resistance are rare (< 1%). If the model simply applied the roof insulation prior probability distribution P(Roof R), shown in Table 4 in the right-hand column, the virtual buildings produced would not correctly represent the actual state of the building stock. This is consistent with the strategy of developing a virtual residential model that is as close to a building stock model as possible, to produce realistic buildings and improve the classification model development process.

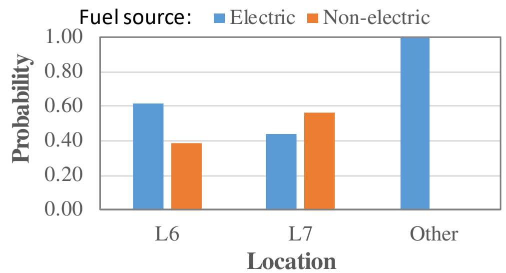
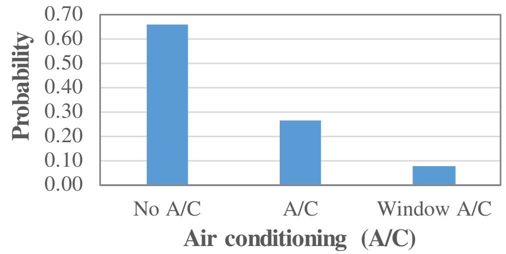
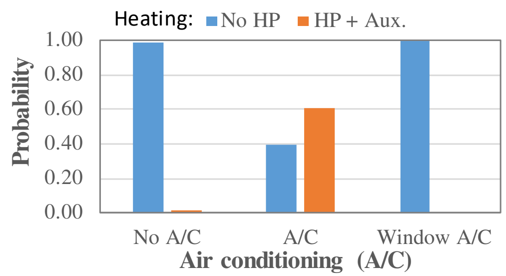
The probability distributions for a number of building parameters are presented in Table 5. Each distribution is described in terms of the number of categories, any dependencies based on other characteristics, a histogram depicting the probability mass function, some notes related to that specific category, and any applicable references for the data source.

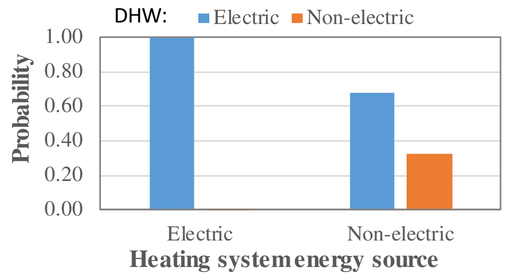
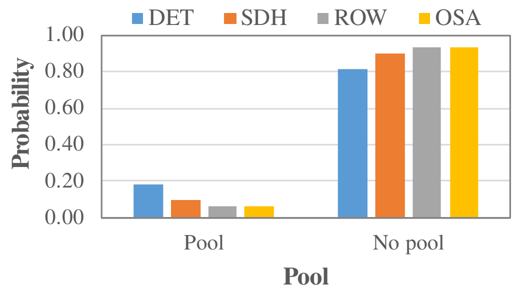
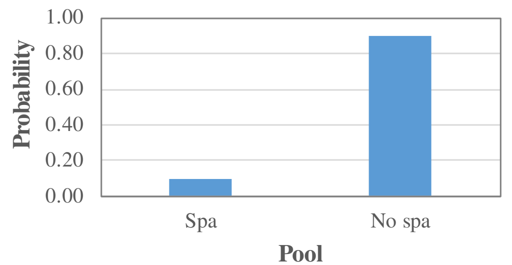
Table 5. Probability distributions for the stochastic parameters generated for each virtual building

<table>
<tr><th>Parameter</th><th>Categories</th><th>Dependencies</th><th>Probability distribution</th><th>Notes</th><th>Reference</th></tr>
<tr><td>Location</td><td>1: Rimouski, Québec (L1)
2: Saguenay, Québec (L2)
3: Québec, Québec (L3)
4: Sherbrooke[i], Québec (L4)
5: Trois-Rivières[i], Québec (L5)
6: Montréal, Québec (L6)
7: Gatineau, Québec (L7)</td><td>N/A</td><td>
</td><td>[i] Also includes several surrounding cities.</td><td>Distribution of total number of SFH in the province of Québec (StatCan 2016; NRCan 2011).</td></tr>
<tr><td>Building type</td><td>1: Single-detached home (DET)
2: Row house (ROW)
3: Semi-detached home (SDH)
4: Other single-attached[i] (OSA)</td><td>Location</td><td>
</td><td>[i] Other single-attached are residential single-family homes adjacent to non-residential buildings, sharing one or more walls.</td><td>Prevalence of each building type by region (StatCan 2016; NRCan 2011).</td></tr>
<tr><td>Number of floors</td><td>1: 1 storey
2: 2 storeys</td><td>Building type</td><td>
</td><td>Probability distributions for ROW, SDH and OSA are based on the same data, i.e. non-single-detached homes.</td><td>Energuide Housing Database (NRCan 2018).</td></tr>
</table>

Parameter	Categories	Dependencies	Probability distribution	Notes	Reference
Occupants	1: 1 occupant 2: 2 occupants 3: 3 occupants 4: 4 occupants 5: 5 occupants[i]	Building type		[i] Includes homes with more than 5 occupants	Statistics Canada household data (StatCan 2011; StatCan 2016). Energuide Housing Database (NRCan 2018).
Heated surface area	1: 56-93 [75][i] m^2 2: 93-139 [115] m^2 3: 139-186 [160] m^2 4: 186-232 [210] m^2 5: >232 [250] m^2	Building type, number of floors		[i] Median value. "Other" includes ROW, SDH and OSA building types. "1" or "2" indicate either one- or two-stories.	Energuide Housing Database (NRCan 2018).
Wall thermal resistance[i]	1: 0.5-1.5 [1.0][ii] m^{2}KW^{-1} 2: 1.5-2.5 [2.0] m^{2}KW^{-1} 3: 2.5- 4.5 [3.0] m^{2}KW^{-1} 5: >4.5 [5.0] m^{2}KW^{-1}	N/A		[i] Total wall assembly thermal resistance [ii] Median value in brackets	Energuide Housing Database (NRCan 2018).

<table>
<tr><th>Parameter</th><th>Categories</th><th>Dependencies</th><th>Probability distribution</th><th>Notes</th><th>Reference</th></tr>
<tr><td>Roof thermal resistance[i]</td><td>1: 0.5-1.5 [1.0][ii] m^2KW^{-1}
2: 1.5-2.5 [2.0] m^2KW^{-1}
3: 2.5-3.5 [3.0] m^2KW^{-1}
4: 3.5-4.5 [4.0] m^2KW^{-1}
5: 4.5-5.5 [5.0] m^2KW^{-1}
8: >5.5 [8.0] m^2KW^{-1}</td><td>Wall thermal resistance</td><td>
</td><td>[i] Total roof assembly thermal resistance
[ii] Median value in brackets</td><td>Energuide Housing Database (NRCan 2018).</td></tr>
<tr><td>Foundation thermal resistance[i]</td><td>1: 0.5-1.5 [1.0][ii] m^2KW^{-1}
2: 1.5-2.5 [2.0] m^2KW^{-1}
3: 2.5-3.5 [3.0] m^2KW^{-1}
4: 3.5-4.5 [4.0] m^2KW^{-1}</td><td>Wall thermal resistance</td><td>
</td><td>[i] Total foundation assembly thermal resistance
[ii] Median value in brackets</td><td>Energuide Housing Database (NRCan 2018).</td></tr>
<tr><td>Equivalent Leakage Area</td><td>1: 248 cm^2 @4Pa
2: 406 cm^2 @4Pa
3: 556 cm^2 @4Pa
4: 775 cm^2 @4Pa
5: 1426 cm^2 @4Pa</td><td>Heated surface area</td><td>
</td><td>N/A</td><td>Energuide Housing Database (NRCan 2018).</td></tr>
</table>

Parameter	Categories	Dependencies	Probability distribution	Notes	Reference
Auxiliary heating	1: Electric 2: Non-electric	Location[i]		[i] Other heating sources are more common in regions L6 and L7 of the study	Energuide Housing Database (NRCan 2018).
Air conditioning	1: No air conditioning 2: Heat pump[i] 3: Window air conditioner			[i] Air-source heat pumps only.	Energuide Housing Database (NRCan 2018).
Heat pump (heating)	1: No heat pump 2: Heat pump[i] + Auxiliary	Air conditioning[ii]		[i] Air-source heat pumps only. [ii] Homes without A/C or with window A/C rarely had heat pumps for heating.	Energuide Housing Database (NRCan 2018).

Parameter	Categories	Dependencies	Probability distribution	Notes	Reference
Domestic hot water	1: Electric element 2: Non-electric	Auxiliary heating[i]		[i] Homes with non-electric heating are more likely to have non-electric hot water heaters	Energuide Housing Database (NRCan 2018).
Swimming pool	1: Swimming pool 2: No swimming pool	Building type		N/A	Real estate database (Realtor.ca 2019). Pool energy calculator (Hydro-Québec 2019a).
Spa	1: Spa 2: No spa	N/A		Limited information was available on spa distribution by building type.	Spa energy calculator (Hydro-Québec 2019b).

The probability distributions presented in Table 5 were developed using a number of sources, most predominantly the Energuide Housing Database (EHD) (NRCan 2018). The EHD consists of over 700 000 homes across Canada that have been audited under a home energy efficiency retrofit program. Probability distributions were established by considering the pre-retrofit characteristics for a subset of roughly 27 000 homes in the Province of Québec. The EHD represented the most detailed information available for many of the building characteristics used in the virtual model.

Cumulative probability distributions were developed from each distribution illustrated in Table 5, which are used within the model to assign properties during random number generation.

Fixed parameters

As with any building simulation, a variety of constant values were used where variable values were not necessary or could not effectively be defined. For example, certain material properties were assigned fixed values consistent with material reference databases, such as for the thermal conductivity of wall insulation or the density of concrete. The impact of using a constant value for such cases was minor compared to other variables in the building simulation. An experienced building simulation modeler would be able to select appropriate values for these types of parameters, and for the purpose of brevity they are not reported here.

Window types

Establishing the typical window types for the province of Québec requires a specific procedure due to the complexity of the available data. In the Energuide Housing Database (EHD) (NRCan 2018) the windows are described in terms of Hot2000 codes (NRCan 2019b), which are six hexadecimal digits corresponding to the number of glazings, window coating, fill type, spacer type, window type, and the frame material of a window.

Each Hot2000 window code is characterized in terms of the overall U-value and solar heat gain coefficients (SHGC). These values are illustrated in a bubble chart to visualize the distribution of the window properties (Figure 5). The k-medoid clustering technique is then applied to the window data set to establish the most representative windows as a function of the number of glazings (single, double and triple). The number of clusters *k* is established iteratively to cover at least 75% of the building stock. Through this method, it is established that 3 single-glazed windows, 15 double-glazed windows, and 6 triple-glazed windows represent the studied building stock. The exact characteristics of the windows are described in Table 6.

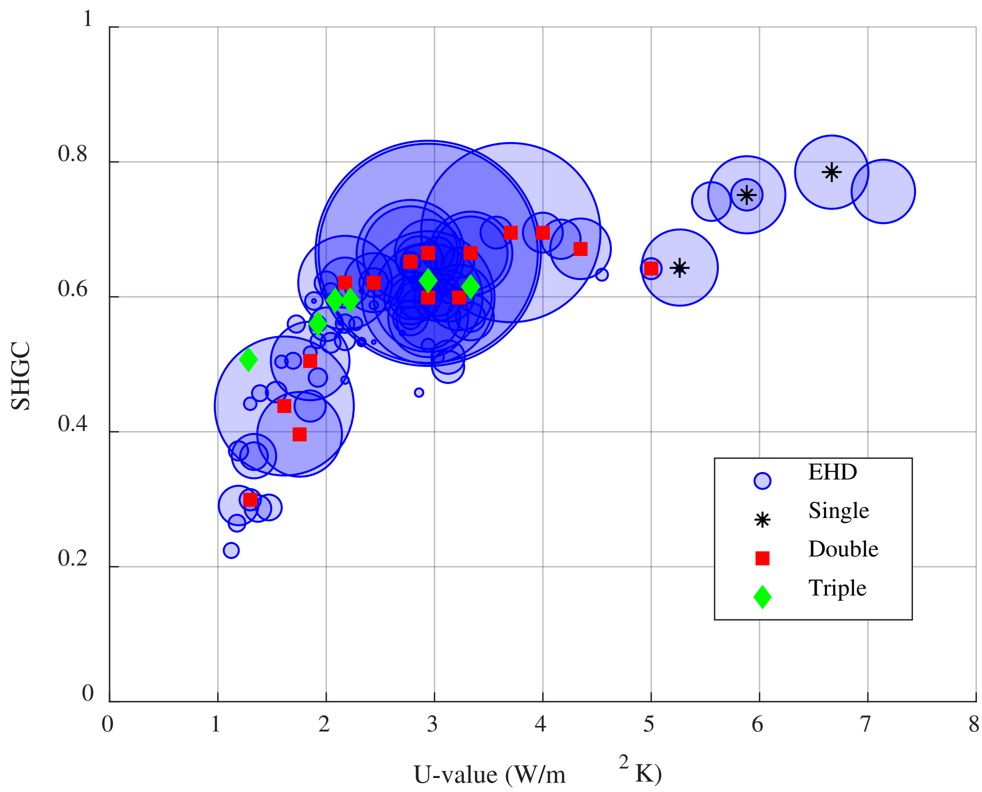


Figure 5. Window data by U-value and SHGC with resulting k-medoid clusters for single-, double- and triple-glazed windows. Size of EHD data points represents number of occurrences in the data set for a given window type.

The probability distribution for each of the 24 window types is established from the number of occurrences of each window type (Table 6). While there are many individual windows reproduced in the model, the objective for window classification is primarily to identify the number of glazings, as opposed to the exact model.

Table 6. Window properties and clustering results

	Window	Probability	Cumulative probability	Hot2000 code	U-value	SHGC	Glazing	Coating	Fill type	Spacer	Type	Frame
Single-glazed windows												
1	**Single 1**	0.020	0.020	100000	6.667	0.785	Single	Clear	13mm air	Metal	Picture	Aluminum
2	**Single 2**	0.026	0.046	100002	5.882	0.751	Single	Clear	13mm air	Metal	Picture	Wood
3	**Single 3**	0.022	0.068	100022	5.263	0.643	Single	Clear	13mm air	Metal	Slider with sash	Wood
Double-glazed windows												
4	**Double 1**	0.002	0.070	200040	5.000	0.642	Double	Clear	13mm air	Metal	Patio door	Aluminum
5	**Double 2**	0.014	0.084	200020	4.348	0.671	Double	Clear	13mm air	Metal	Slider with sash	Aluminum
6	**Double 3**	0.006	0.090	202000	4.000	0.695	Double	Clear	6 mm air	Metal	Picture	Aluminum
7	**Double 4**	0.123	0.212	200000	3.704	0.695	Double	Clear	13mm air	Metal	Picture	Aluminum
8	**Double 5**	0.047	0.259	202002	3.333	0.665	Double	Clear	6 mm air	Metal	Picture	Wood
9	**Double 6**	0.016	0.276	202012	3.226	0.599	Double	Clear	6 mm air	Metal	Hinged	Wood
10	**Double 7**	0.394	0.670	200002	2.941	0.665	Double	Clear	13mm air	Metal	Picture	Wood
11	**Double 8**	0.112	0.782	200012	2.941	0.599	Double	Clear	13mm air	Metal	Hinged	Wood
12	**Double 9**	0.002	0.784	200006	2.778	0.652	Double	Clear	13mm air	Metal	Picture	Fiberglass
13	**Double 10**	0.016	0.800	231002	2.439	0.621	Double	Low-e .20 (hard 1)	9 mm air	Metal	Picture	Wood
14	**Double 11**	0.043	0.843	234002	2.174	0.621	Double	Low-e .20 (hard 1)	9mm argon	Metal	Picture	Wood
15	**Double 12**	0.024	0.867	223214	1.852	0.505	Double	Low-e .10 (soft)	13mm argon	Insulating	Hinged	Vinyl
16	**Double 13**	0.027	0.894	213214	1.754	0.396	Double	Low-e .04 (soft)	13mm argon	Insulating	Hinged	Vinyl
17	**Double 14**	0.074	0.968	213204	1.613	0.438	Double	Low-e .04 (soft)	13mm argon	Insulating	Picture	Vinyl
18	**Double 15**	0.002	0.969	644204	1.299	0.299	Double - 1 heat mirror	Low-e .35 (hard 2)	9mm argon	Insulating	Picture	Vinyl
Triple-glazed windows												
19	**Triple 1**	0.002	0.971	300010	3.333	0.615	Triple	Clear	13mm air	Metal	Hinged	Aluminum
20	**Triple 2**	0.007	0.978	301000	2.941	0.624	Triple	Clear	9 mm air	Metal	Picture	Aluminum
21	**Triple 3**	0.006	0.984	301002	2.222	0.595	Triple	Clear	9 mm air	Metal	Picture	Wood
22	**Triple 4**	0.005	0.989	300002	2.083	0.595	Triple	Clear	13mm air	Metal	Picture	Wood
23	**Triple 5**	0.008	0.997	331002	1.923	0.560	Triple	Low-e .20 (hard 1)	9 mm air	Metal	Picture	Wood
24	**Triple 6**	0.003	1	323204	1.282	0.507	Triple	Low-e .10 (soft)	13mm argon	Insulating	Picture	Vinyl

Climate files

The location of the home determines the climate file used in the building simulation. For the purpose of generating realistic electricity consumption values via building simulation, weather data files are used for the year 2016. Classification based on location will therefore be able to more consistently determine the location of an anonymous smart meter data based on the order of magnitude of the electricity load in a heating-dominated climate. Other climate data can be used to generate the VSM data sets by substituting the regional climate files for other years.

VSM Data

The overall procedure to produce the VSM data set is depicted in Figure 2, which shows that the ultimate goal is to produce electricity smart meter data with known building characteristics. A data set of 200 000 VSM profiles is provided with this paper, which consists of input data, VSM data, load profiles, and annual totals for heating, cooling, lighting, equipment and domestic hot water electricity.

Input data

A virtual smart meter profile is produced by first generating a single-family home. Each home is characterized by the uniform probability distributions (UPD) and probability mass functions (PMF) mentioned previously. Each distribution requires the generation of a distinct random number, which is then used to determine which value to use for a given parameter. A sample input set is presented in Table 7 as an example of the link between the random number generated for each parameter and the corresponding value used in the simulation.

Table 7. Sample input set based on random number generation

Parameter	**Random number**	**Bin #**	**Total # bins**	**Distribution type**	**Corresponding value**
Location	0.554	6	7	PMF	Montréal, Canada
Building type	0.536	1	4	PMF	Detached
Occupancy profile number	0.350	06[1]	15	UPD	Profile #6
Window # glazings	0.961	2	3	PMF	Double-glazed windows
Surface area	0.403	3	5	PMF	160 m^2
Window-to-wall ratio	0.120	1	3	UPD	0.1
Building rotation	0.162	1	4	UPD	0° rotation
Occupants	0.117	1	5	PMF	1 occupant
Building adjacency	0.561	1	4	UPD	Detached - no adjacency
Floors	0.131	1	2	PMF	1 floor
Wall R[2]	0.441	2	4	PMF	2 m^2KW^{-1}
Roof R	0.110	2	6	PMF	2 m^2KW^{-1}
Foundation R	0.901	2	4	PMF	2 m^2KW^{-1}
Infiltration rate	0.976	5	5	PMF	1426 cm^2
Air conditioning	0.594	1	3	PMF	No AC
Heat pump	0.272	1	2	PMF	No heat pump
Auxiliary heating type	0.809	1	2	PMF	Electric
DHW type	0.035	1	2	PMF	Electric
Aspect ratio	0.391	2	5	UPD	0.9
Pool	0.242	1	2	PMF	No pool
Spa	0.745	1	2	PMF	No Spa

[1] Requires two digits due to 15 possible bins.
[2] R: Thermal resistance.

The resulting bin numbers from the example above correspond to a unique combination of input values for a particular electricity consumption data profile. The building characteristics can therefore be traced back for each VSM data profile, to the nearest bin value. In some cases the exact value used in the building simulation is also provided, such as for the heated surface area. In the case that a duplicate input set exists, the duplicate is removed in postprocessing and a new profile is generated and added to the data set.

VSM profile data

The total house electricity consumption in kilowatt-hours (kWh) is recorded at 15-minute intervals for 365 days. This accounts for 35 040 data points per VSM profile, not including input data, which is stored separately. The data set with 200 000 virtual homes requires approximately 50 gigabytes of storage space in an uncompressed format.

Load profiles

The internal load profiles generated with the CREST tool are provided with the VSM data for reference. Each profile is based on the number of occupants (from 1 to 5) and a randomly generated profile (from 1 to 15), resulting in 75 different possible load profiles. Lighting, appliance and domestic hot water loads are based on the occupancy profiles. A user can refer to the *Occupancy Profile Number* and *Occupants* inputs to determine the corresponding load profile that was used in the building simulation. Data is organized in terms of the following characteristics:

- Occupants in the home
- Occupants active
- Lighting energy use
- Appliance energy use
- Domestic hot water energy use

Annual totals for heating, cooling, lighting, equipment and domestic hot water electricity use

In addition to the smart meter data at 15-minute intervals, annual total electricity use for heating, cooling, lighting, equipment and domestic hot water is also provided. This allows for a greater understanding of the electricity consumption within a home without adding a significant amount of data to an already very large data set. Annual totals are provided in kilowatt-hours (kWh).

Overview of the data

The generated homes for the VSM profile framework are intended to cover a range of buildings that represent the most likely combinations of parameters. A box and whisker plot of the annual electricity consumption for the virtual smart meter data set is shown in Figure 6, which illustrates the range of values for heating, cooling, electricity, equipment, domestic hot water (DHW), and total electricity consumption. The box represents the data that lies between the 25th and 75th percentiles for the data set, i.e. the interquartile range (IQR). The whiskers represent values within 1.5 times the IQR. Data outside the whiskers are considered outliers and marked with individual data points. The median value is indicated by a red line within the box plot. The virtual profiles are compared to the average single-family home electricity consumption for the province of Québec (NRCan 2019).

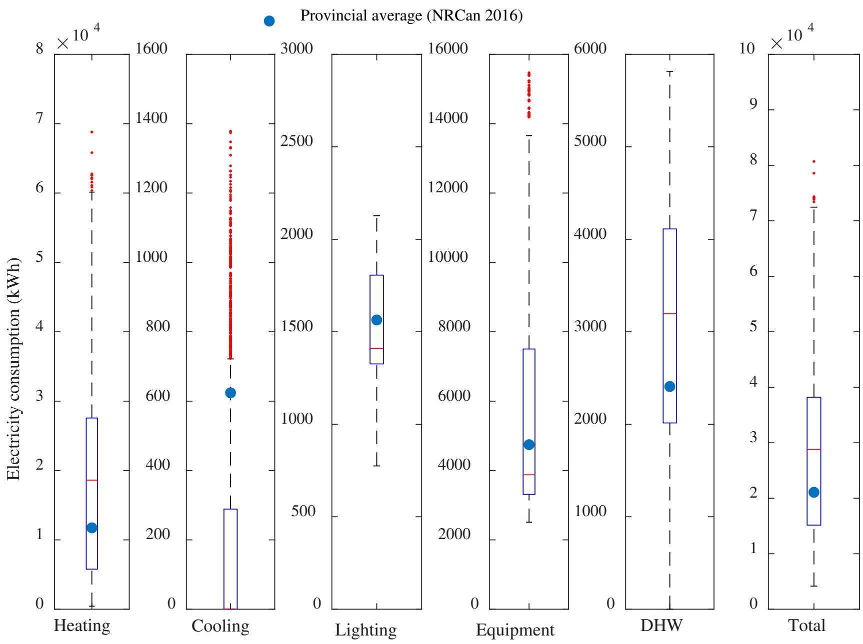
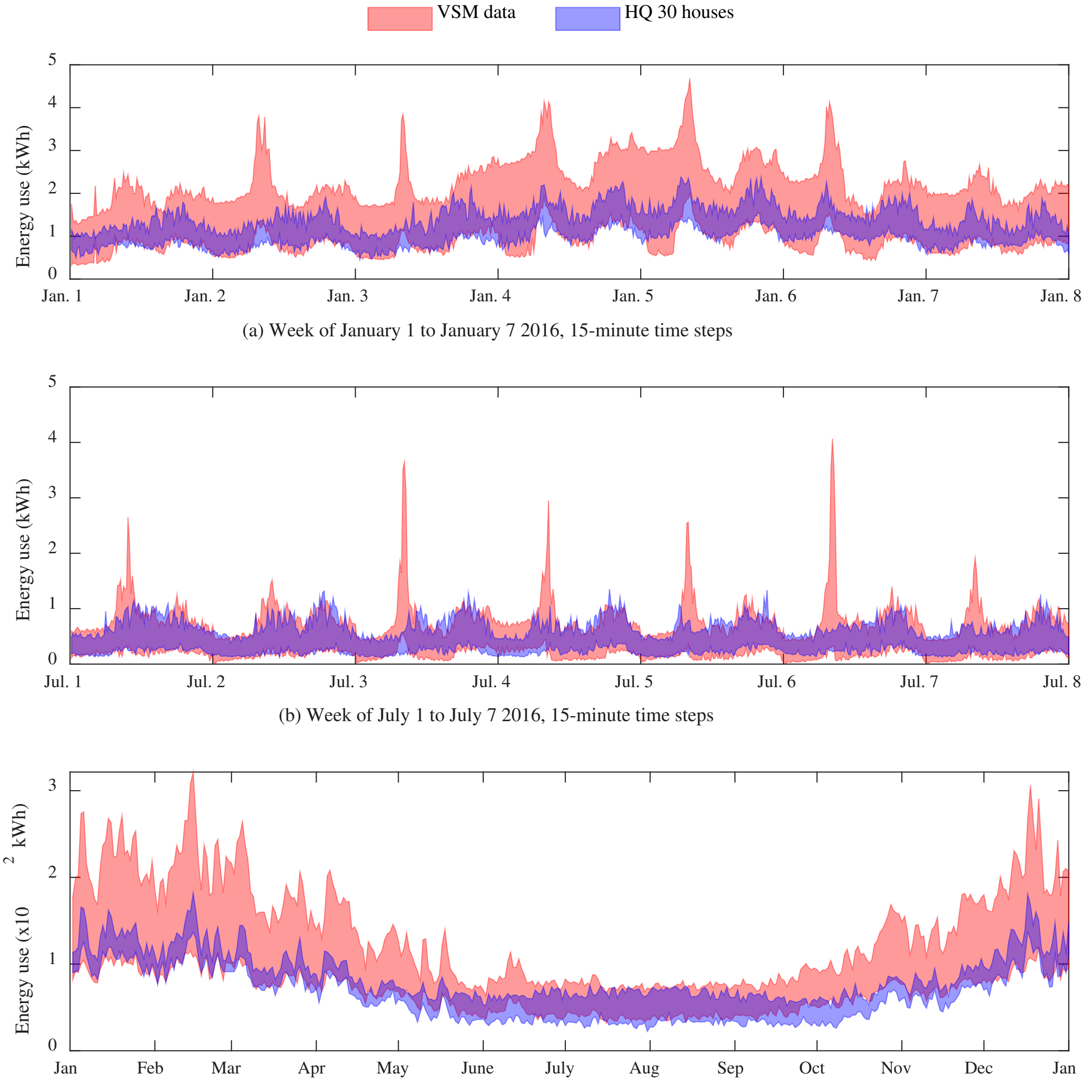


Figure 6. Box and whisker plot illustrating the variation in electricity consumption when compared with provincial averages

The virtual homes vary quite widely in terms of electricity consumption, which is consistent with the variety of building parameters used to produce the data set. Buildings with non-electric heating would have negligible heating electricity consumption, while large, poorly insulated homes with large families would have a relatively high overall electricity consumption. Cooling values tend to be underestimated by the model, which is likely due to the underlying assumptions used in the VSM framework that underestimate the number of residences equipped with air conditioning systems in the studied building stock. Equipment values vary widely depending on the presence of a pool and/or spa in the house. Overall the total electricity consumption is consistent with provincial averages for Québec.



(a) Week of January 1 to January 7 2016, 15-minute time steps

(b) Week of July 1 to July 7 2016, 15-minute time steps

(c) Annual electricity for 2016, daily time steps

Figure 7. Interquartile ranges for the VSM data set and 30 houses in the province of Québec for (a) the first week of January, (b) the first week of July, and (c) daily energy use for a full year.

To further illustrate the range of possible profiles in the virtual smart meter data set, the complete VSM data for all locations is compared to the measured electricity consumption of 30 houses in location #5 (see Table 5) in Figure 7. Each graph depicts the interquartile range (25th to 75th percentiles) for the respective data sets at each time step. The results in Figure 7(a) and Figure 7(b) illustrate the quarter-hourly electricity consumption for a week in winter and a week in summer, respectively. Daily energy use is compared for a full year in Figure 7(c). For all three cases, the measured smart meter data generally falls within the range of values produced in the virtual smart meter data set. Peak energy use tends to be higher in the virtual profiles when compared with the measured data indicating that there are some houses in the VSM data set that are poorly insulated compared to the 30 houses which tend to be better built. Nonetheless, the results in Figure 7 indicate that the virtual smart meter data includes a range of cases consistent with the measured data. In the future, a larger measured data set will be used to compare with the VSM data.

Discussion

The Virtual Smart Meter data set is provided in a format with sufficient accompanying information to be useful for a variety of purposes. By considering the input data, load profiles and corresponding smart meter data, researchers can utilize the data set to verify their own models, study electricity consumption for a variety of housing types, or estimate electricity consumption for a large number of houses, among other possibilities. In this discussion, some commentary on the general methodology for developing a virtual smart meter data set for another building stock is provided. Two example applications for the virtual smart meter data set are then provided: 1) developing inverse models using classification of electricity smart meter data, and 2) verification of load disaggregation algorithms.

Developing and using a virtual smart meter data set

The methodology presented in this paper can be applied towards developing a virtual smart meter data set for a single-family home residential building stock. Suggestions have been made to aid in seeking out sources of information on building and occupant characteristics. The segmentation and characterization processes of building stock modeling are necessary to develop probability distributions for the parameters used in the VSM building generator. These distributions can be used to develop a Bayesian network, such as the one illustrated in Figure 4, to produce combinations of building inputs that actually match the selected building stock. Less common or impossible combinations of building parameters should not be prioritized when producing a virtual smart meter data set. Alternatively, less categories can be used to characterize the building parameters and reduce the overall number of combinations of inputs.

When producing virtual smart meter profiles, parallel processing is an essential component to reducing the overall time required to complete the virtual data set. Cloud computing can allow for multiple building simulations in parallel. In the case of this study, 20 simulations were run in parallel, with each annual simulation requiring approximately 15 seconds. A full data set of 200 000 profiles requires approximately 42 hours to produce under these conditions.

Working with large smart meter data sets can require significant computational resources, whether they are virtual or measured data. The main limitation is in the random access memory (RAM) for loading and processing a significant number of smart meter profiles (>100 000), which scales based on the complexity of the building stock and the desired accuracy. This issue can be offset by aggregating the electricity consumption to different time scales, such as hourly or daily data, to significantly reduce the required memory use. Alternatively, studying specific periods of time, such as a single week of data at 15-minute intervals, can reduce the memory requirements to manageable levels. Finally, filtering the smart meter data based on specific building parameters can also be a viable option.

Classification modeling of electricity smart meter data

As described in the literature review, few public smart meter data sets contain detailed information about the buildings. The data sets that exist have limited information about the houses or are too narrow in the quantity of buildings studied, which restricts the range of parameters studied. As an example, in order to evaluate the impact of the building envelope on the electricity meter data, detailed envelope performance data is required. If a classification modeler wishes to estimate the level of thermal resistance for a home based on the smart meter data, a smart meter data set with known envelope properties is required to train a classification model. The VSM data set can be used for this purpose.

For an example of this method applied in practice, readers can refer to the study by Neale et al. (2019), where linear discriminant analysis was applied to a preliminary version of the VSM data set to predict a number of building parameters for real smart meter data. The classification process significantly increased the accuracy of predicting building parameters when compared to random guessing. Computational resources for classification are also discussed in Neale et al. (2019).

Automated load disaggregation algorithms

The virtual smart meter data set can be used to verify the effectiveness of automated electricity load disaggregation methods, which are commonly used to divide electricity smart meter data into heating, appliance, lighting, and other relevant loads when only aggregated data is available. Such methods are commonly applied to limited data sets, such as those described in Table 1, for which there is little variety in the appliance, occupancy and load profiles to test the algorithms. Deb et al. (2019) developed a load disaggregation algorithm for electric heating and tested it on a data set for a single home with 37 days of data. While the house was well-parametrized, the extent of the validation was limited by the scope of the data set.

The VSM data set can be used to test load disaggregation algorithms for a variety of building geometry, occupant behaviour, appliance and lighting configurations and other factors. Researchers can apply their algorithms to the virtual smart meter data and compare it directly to the submetered heating, cooling, lighting, appliance and domestic hot water subtotals. The impact of each building parameter can be studied in order to improve the accuracy of the tested methods.

Conclusion

A virtual smart meter (VSM) data set is an effective tool for evaluating the impact that building parameters have on electricity consumption. The provided data set and methodology can serve as an example for other researchers on producing and structuring VSM data for other building stocks. In addition, there are many possible applications that require smart meter data with known building characteristics, including load disaggregation algorithms, classification modeling, peak load studies, technology evaluations and other work. The smart meter data sets that currently exist in literature limit the effectiveness of studies in these fields. In addition, the use of a model-based smart meter data set allows the user to filter the data based on specific inputs to fit the desired approach. The provided VSM data set can be used by researchers to verify their methods and provide insight on residential electricity consumption.

The VSM framework and data sets will be improved in a number of ways in the future. The authors intend to continue to develop the Bayesian network defining the dependencies between the building characteristics. While the VSM framework is not intended to be a building stock model at this stage of the work, eventually it is the hope of the authors to improve the framework to the point where the produced data set is as representative as possible of the building stock, improving upon the results in Figure 6. In addition, as additional information is obtained on the building stock the probability distributions for each parameter will be updated. Work is also ongoing to extract building data from smart meter data using classification modeling, which will provide an additional source of information on the building stock.

Acknowledgements

The authors gratefully acknowledge the financial support of an IVADO Fundamental Research Grant [PRF-2017-12].

Data set

A data set of 200,000 virtual smart meter profiles with corresponding building characteristics can be found at http://vsmdata.meca.polymtl.ca/.

References

Akshay Uttama Nambi, S. N., Antonio Reyes Lua, and R. Venkatesha Prasad. 2015. "LocED: Location-Aware Energy Disaggregation Framework." In *BuildSys 2015 - 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, 45–54. Association for Computing Machinery, Inc. doi:10.1145/2821650.2821659.

ASHRAE. 2013. *ASHRAE Handbook of Fundamentals*. Altanta, GA: ASHRAE.

Barker, Sean, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. 2012. "Smart*: An Open Data Set and Tools for Enabling Research in Sustainable Homes." In *Proc. SustKDD '12*, 1–6.

Beckel, Christian, Leyna Sadamori, Thorsten Staake, and Silvia Santini. 2014. "Revealing Household Characteristics from Smart Meter Data." *Energy* 78: 397–410.

Booth, A.T., R. Choudhary, and D.J. Spiegelhalter. 2012. "Handling Uncertainty in Housing Stock Models." *Building and Environment* 48 (February). Pergamon: 35–47. doi:10.1016/J.BUILDENV.2011.08.016.

Carroll, Paula, Tadhg Murphy, Michael Hanley, Daniel Dempsey, and John Dunne. 2018. "Household Classification Using Smart Meter Data." *Journal of Official*

Statistics 34 (1): 1–25.

CER. 2012. *CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010*. 1st ed. Irish Social Science Data Archive. SN: 0012-00.

Deb, Chirag, Mario Frei, Johannes Hofer, and Arno Schlueter. 2019. "Automated Load Disaggregation for Residences with Electrical Resistance Heating." *Energy and Buildings* 182 (January). Elsevier Ltd: 61–74. doi:10.1016/j.enbuild.2018.10.011.

Hammerstrom, D J, R Ambrosio, J Brous, T A Carlon, D P Chassin, J G Desteese, R T Guttromson, et al. 2007. *Pacific Northwest GridWise^{TM} Testbed Demonstration Projects; Part I. Olympic Peninsula Project*.

Hammerstrom, D J, J Brous, T A Carlon, D P Chassin, C Eustis, G R Horst, O M Järvegren, et al. 2007. *Pacific Northwest GridWise^{TM} Testbed Demonstration Projects; Part II. Grid Friendly^{TM} Appliance Project*.

Hydro-Québec. 2012. *Réponse d'Hydro-Québec Distribution Aux Engagements 3, 10, 15, 18, 21 (UC), 21 (UMQ) ET 22 À 27*.

Hydro-Québec. 2016. *Rapport Annuel 2015*. Montréal, Canada.

Hydro-Québec. 2019a. "Pool Energy Calculator." http://www.hydroquebec.com/residential/customer-space/electricity-use/tools/swimming-pool-calculator.html.

Hydro-Québec. 2019b. "Spa Energy Calculator." http://www.hydroquebec.com/residential/customer-space/electricity-use/tools/spa-calculator.html.

IEA (International Energy Agency). 2017. *Energy Efficiency 2017*.

IEA (International Energy Agency). 2019. "Smart Grids - Tracking Clean Energy Progress." https://www.iea.org/tcep/energyintegration/smartgrids/.

Johnson, Geoffrey, and Ian Beausoleil-Morrison. 2017. "Electrical-End-Use Data from

23 Houses Sampled Each Minute for Simulating Micro-Generation Systems." *Applied Thermal Engineering* 114 (March). Pergamon: 1449–1456. doi:10.1016/J.APPLTHERMALENG.2016.07.133.

Kleiminger, Wilhelm, Christian Beckel, and Silvia Santini. 2015. "Household Occupancy Monitoring Using Electricity Meters." In *2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015)*, 975–986. doi:10.1145/2750858.2807538.

Klein, S.A., W.A. Beckman, J.W. Mitchell, J.A. Duffie, N.A. Duffie, T.L. Freeman, J.C. Mitchell, et al. 2017. "TRNSYS 18: A Transient System Simulation Program, Solar Energy Laboratory, University of Wisconsin, Madison, USA." http://sel.me.wisc.edu/trnsys.

Makonin, Stephen, Bradley Ellert, Ivan V. Bajić, and Fred Popowich. 2016. "Electricity, Water, and Natural Gas Consumption of a Residential House in Canada from 2012 to 2014." *Scientific Data* 3 (160037): 1–12. doi:10.1038/sdata.2016.37.

Mathworks. 2018. "Matlab R2018b."

McKenna, Eoghan, and Murray Thomson. 2016. "High-Resolution Stochastic Integrated Thermal–electrical Domestic Demand Model." *Applied Energy* 165: 445–461.

Murray, David, Lina Stankovic, and Vladimir Stankovic. 2017. "An Electrical Load Measurements Dataset of United Kingdom Households from a Two-Year Longitudinal Study Background & Summary." *Nature - Scientific Data*. doi:10.1038/sdata.2016.122.

Neale, Adam, Michaël Kummert, and Michel Bernier. 2018. "Generator : A Stochastic Virtual Smart Meter Data Generation Model for Residential Building Stock

Characterization.” In *ESim 2018, the 10th Conference of IBPSA-Canada*, 65–74. Montreal, Canada.

Neale, Adam, Michaël Kummert, and Michel Bernier. 2019. “Linear Discriminant Analysis for Classification of Building Parameters for a Large Virtual Smart Meter Data Set.” In *Proceedings of the 16th IBPSA Conference*, 8. Rome, Italy.

NRCan (Natural Resources Canada). 2011. *Survey of Household Energy Use*.

NRCan (Natural Resources Canada). 2018. “Energuide Housing Database.”

NRCan (Natural Resources Canada). 2019a. “Residential Sector - Table 1: Secondary Energy Use and GHG Emissions by Energy Source.” *National Energy Use Database*. http://oee.nrcan.gc.ca/corporate/statistics/neud/dpa/showTable.cfm?type=CP&sector=res&juris=qc&rn=1&page=0.

NRCan (Natural Resources Canada). 2019b. “Hot2000.” https://www.nrcan.gc.ca/energy-efficiency/energy-efficiency-homes/professional-opportunities/tools-industry-professionals/20596.

NRCan (Natural Resources Canada). 2020. “Canada’s Secondary Energy Use (Final Demand) by Sector, End Use and Subsector.” http://oee.nrcan.gc.ca/corporate/statistics/neud/dpa/showTable.cfm?type=HB&sector=aaa&juris=ca&rn=2&page=0.

Realtor.ca. 2019. “Comprehensive Real Estate Listing Database.” https://www.realtor.ca/.

Reinhart, Christoph F., and Carlos Cerezo Davila. 2016. “Urban Building Energy Modeling – A Review of a Nascent Field.” *Building and Environment* 97: 196–202.

Saldanha, Neil, and Ian Beausoleil-Morrison. 2012. “Measured End-Use Electric Load

Profiles for 12 Canadian Houses at High Temporal Resolution." *Energy and Buildings* 49 (June). Elsevier: 519–530. doi:10.1016/J.ENBUILD.2012.02.050.

Sherman, M.H., and D.T. Grimsrud. 1980. "Infiltration-Pressurization Correlation: Simplied Physical Modeling." *ASHRAE Transactions* 86 (2): 778.

StatCan. 2011. "Household Size, by Province and Territory." *Census of Population and Statistics Canada Catalogue No. 98-313-XCB*.

StatCan. 2016. "2016 Census." *Statistics Canada*. http://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm.

Swan, Lukas G., and V. Ismet Ugursal. 2009. "Modeling of End-Use Energy Consumption in the Residential Sector: A Review of Modeling Techniques." *Renewable and Sustainable Energy Reviews* 13 (8): 1819–1835.