Discriminant Analysis Classification of Residential Electricity Smart Meter Data

Adam Neale*, Michaël Kummert and Michel Bernier

Département de Génie Mécanique, Polytechnique Montréal, Montreal, Canada;

Corresponding author: Adam Neale, adam.neale@polymtl.ca

Accepté pour publication dans Energy and Buildings – 27 décembre 2021

Discriminant Analysis Classification of Residential Electricity Smart Meter Data

The objective of this study is to apply machine learning classification to predict building characteristics from electricity smart meter data for the purpose of building stock characterization. Given that there are no publicly available large-scale residential electric smart meter data sets with detailed building characteristics, an open-source virtual smart meter (VSM) data set is used. The VSM data consists of electricity consumption profiles for 200,000 homes with 21 known characteristics, which are used to train predictive models with linear discriminant analysis (LDA). The classification accuracy (CA) is determined for a variety of scenarios where the meter data aggregation and period are varied. The CA depends on the parameter to be classified (the *class*), the number of data points per building (the *features*) and the number of buildings used for classification. Reliable classification results are obtained when the number of buildings exceeds the number of features by a significant margin. An application of the developed predictive models to a small data set of 30 real houses illustrates the usefulness of the method but also the challenges in achieving a generalized model with virtual data.

1 Introduction

Evaluating the effectiveness of energy efficiency measures and technology upgrades for buildings on a large scale, such as at the urban, provincial or national levels, can require the use of a building stock energy model. Developing such models can be accomplished using a number of techniques, including top-down models and bottom-up engineering and statistics-based models (Swan and Ugursal 2009). Building archetypes are one such method that requires an information gathering process known as segmentation and characterisation (Sokol et al. 2016). Regardless of the technique used, information on the building stock is a limiting factor on the accuracy of the resulting model (Booth, Choudhary, and Spiegelhalter 2012).

Electricity smart metering has become very widespread in the last decade, with the United States installing 98 million meters in 2019, the total now covering 70% of the U.S. residential market (Mordor Intelligence 2021). In the province of Québec, Canada, there are 3.7 million installed smart meters, which includes over 1 million single-family homes (Hydro-Québec 2016). The prevalence of metered data can provide a wealth of information considering that the electricity consumption is monitored at subhourly intervals. With a sufficiently large smart meter data set, with information about relevant building parameters, it could be possible to leverage the vast quantity of metered data to extract building details from anonymous smart meter data. Such details could be used for the purpose of building stock modelling techniques, such as building stock segmentation and characterisation (Sokol et al. 2016). The main problem is to leverage such smart meter data in a way that protects the privacy of the homeowners while serving modeling professionals as a source of building stock data.

A variety of well-documented supervised machine learning techniques can serve to establish correlations between a set of inputs and one or more data points. For example, James et al. (2013) provides an introduction and overview of many statistical learning techniques. The term *supervised* indicates that a training set is required to develop a model that correlates a series of input data, or *predictors*, with a corresponding output, or *response*. This process is also often called *classification* when the response is qualitative rather than quantitative, as the process determines a *class* category for a particular set of data, as opposed to a numeric value. A generalized illustration of predictive model development and application is described in Figure 1.

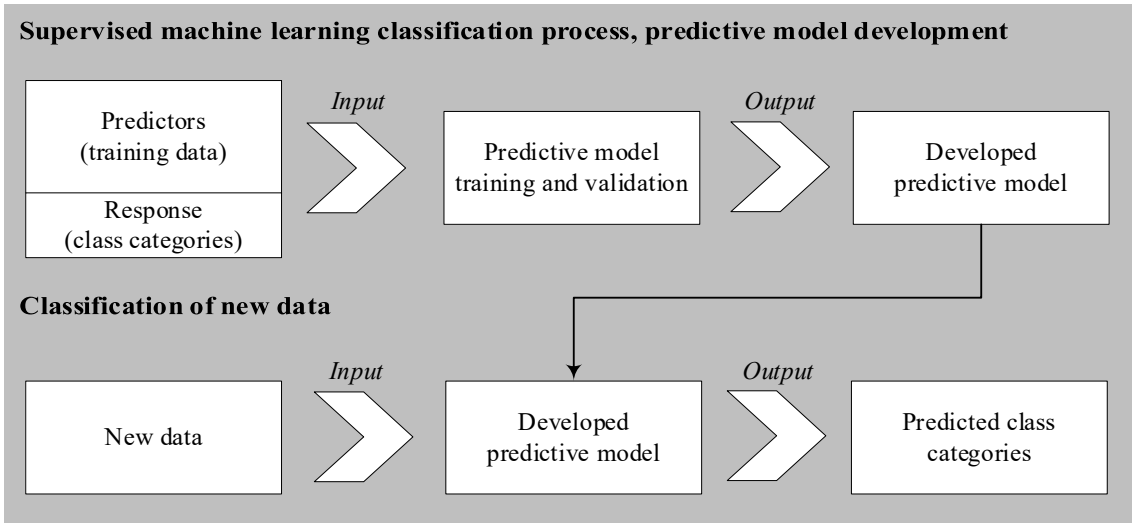

Figure 1. Generalized supervised machine learning predictive model development process

The general approach to develop a predictive model involves a training set of predictor data used to explain a qualitative response variable described in terms of class categories, as described in Figure 1. The term *predictive model* is used here, as the purpose of classification is to develop a model such that new data can be input to predict a class value (Shmueli 2010). The new predictor data must share the same number of features as the original training data set, which describes the number of data for each case. The classification terms used above, such as predictor, response and feature, are described in more detail in the glossary at the end of the paper.

Statistical learning methods, including the classification approach illustrated in Figure 1, are already applied to a wide variety of fields, including handwriting recognition, DNA mapping, e-mail spam detection, etc. (Hastie, Tibshirani, and Friedman 2009). Classification has been identified as one of the key smart grid analysis tools going forward (Y. Zhang, Huang, and Bompard 2018). Supervised machine learning classification could be used on smart meter data provided some information is known about the buildings in the data set, such as the surface area, location, etc. However, studies have shown there are very few residential smart meter data sets with sufficient information about the houses (Neale, Kummert, and Bernier 2020a).

The objective of this study is to leverage a virtual smart meter data set in order to develop predictive models to predict building characteristics from real electricity smart meter data and ultimately improve the building stock characterization process. Linear discriminant analysis is evaluated as a technique to perform this task. Predictive models are developed for 21 known building parameters using the virtual data set of 200,000 buildings. The influence of the data set size and the feature selection on the classification accuracy is presented. The developed models are applied using a set of 30 houses with real smart meter data with known building parameters to test the generalization of the predictive models. More specifically, this paper aims to:

1. Demonstrate how LDA can be used to classify electricity smart meter data, using a set of virtual smart meter data designed for that purpose;
2. Illustrate the impact of the data set size and number of features on the classification accuracy;
3. Guide future users of LDA on potential problems when developing predictive models on large data sets, with feature selection recommendations for specific building parameters;
4. Demonstrate the limitations of applying real smart meter data to a predictive model developed using virtual data.
5. Present how this approach can be applied for building stock energy model development, a current need in industry.

This paper builds upon a previous work (Neale, Kummert, and Bernier 2019) which laid the groundwork for the present study. This journal paper contains entirely new results, significantly increased detail on the methods, new forms of presentation of the results, additional analysis and conclusions to guide those wishing to use LDA classification.

2 Literature review

Data required for housing stock model development is one of the key limiting factors for accurate stock energy prediction (Booth, Choudhary, and Spiegelhalter 2012). Smart meter data presents a potential untapped opportunity for insight into every residential building, but due to privacy reasons it is most often anonymous and without any information on the building's characteristics. A building's parameters could be predicted using classification with a sufficient training set, but to the authors' knowledge no studies have fully evaluated the potential of doing so.

Following are some works describing common smart meter data analytics applications and techniques. The focus of the first portion of the review is to describe common methods and practices related to machine learning in building applications, followed by smart meter data analytics with a few examples. Next, a review targeting previous works in supervised machine learning of smart meter data for the purpose of predicting building parameters is performed. While the focus of the authors is residential energy consumption, where applicable non-residential cases are examined as well.

2.1 Machine learning in building applications

Machine learning (ML) can be categorized as supervised ML, unsupervised ML, semi-supervised ML and reinforcement learning techniques. Sarker (2021) provides a comprehensive review of machine learning techniques with descriptions and general applications. Supervised learning, which is the focus of this study, can be divided into two categories depending on whether the studied variables are discrete or continuous. Machine learning on discrete variables is referred to as *classification*, while for continuous variables it is known as *regression*. There are many reference texts

describing the statistical derivation of methods in machine learning as well, such as the one by Hastie et al. (2009).

Machine learning has become commonplace for a variety of building applications. Djenouri et al. (2019) provide an overview of ML in smart building applications, which summarizes a wide variety of statistical methods that are divided in two broad categories: occupant-centric and energy/device centric applications. Occupant-centric machine learning focuses on occupancy detection, activity recognition and preference/behaviour identification. Energy/device-centric applications include energy profiling and demand estimation, appliance profiling and fault detection, and sensor inference.

The reviews by Sarker (2021) and Djenouri et al. (2019) provide comprehensive descriptions of a variety of ML methods and algorithms. Some specific examples of ML applications in buildings are provided here as well. For example, Gładyszewska-Fiedoruk and Sulewska (2020) applied linear discriminant analysis (LDA) classification and artificial neural networks (ANN) on thermal comfort surveys to evaluate occupant responses to various building indoor environmental conditions. Esen et al. (2008) use ANN and adaptive neuro-fuzzy inference systems (ANFIS) to forecast the performance of ground-source heat pumps under a variety of conditions. Li et al. (2016) apply LDA to perform fault detection and diagnosis (FDD) on a chiller, which demonstrated the effectiveness of multiscale classification for FDD of mechanical systems in buildings. These studies illustrate how the use of ML has permeated many facets of the field of building engineering, while the focus of the authors is specifically on smart meter data analysis.

2.2 Smart meter analytics

Wang et al. (2018) performed a thorough review of smart meter data analytics methods.

Applications identified include load analysis, load forecasting, load management and other various subcategories. Techniques include time series analysis, dimension reduction, outlier detection, classification, clustering, deep learning, and more. While Wang identifies classification as a relevant technique, building characterization is not listed in the review. Few cases have been found in the literature of supervised machine learning on smart meter data for residential building characterisation. Many works have used regression and clustering techniques on thermal and electricity metered data, both supervised and unsupervised. Many works are cited by Wang et al. (2018) for interested readers, and a few examples are provided here for context.

Classification and other machine learning (ML) algorithms have been applied to smart meter data in recent works, but in many cases in a context of anomaly detection (L. Zhang et al. 2019; Himeur et al. 2021; Oprea et al. 2021). These works aim to detect unusual meter data that may affect energy analysis techniques, such as load forecasting and/or profiling, as well as energy theft detection. ML has been applied in specific smart meter applications, such as identifying changes in occupant behaviour via metered data pattern recognition, for the purpose of diagnosing at-risk patients in distress (Chalmers et al. 2019). Non-intrusive load monitoring (NILM) is another frequent application where classification and supervised machine learning are applied (Klemenjak 2018).

Gianniou et al. (2018) applied regression techniques to daily thermal energy data to predict temperature setpoint and building envelope characteristics for 14,000 houses in Denmark. While classification was not used in their study, properties of a building stock are successfully extracted from meter data with some degree of accuracy. The study is limited by the information available on each building, as only the weather,

heating energy and basic building geometry were available to develop linear regression models.

Unsupervised support-vector regression was applied by Westermann et al. (2020) on electricity smart meter data to predict heating systems for two sets of 400 buildings. Clustering techniques were applied to identify different energy signatures from metered data to identify heating system types. For a case study applied to British Columbia, Canada, the authors were able to accurately predict the distribution of heating systems corresponding to the provincial average, within 2% per type. However, no actual system data was available to validate the prediction at the building-level.

Ullah et al. (2020) applied deep learning clustering techniques to monthly residential building stock energy data for the purpose of identifying energy consumption patterns. The work identifies clusters of energy consumption levels in stock data, and also analyzes a single house's energy consumption over several years. Self-organizing maps are employed to cluster the data after a detailed encoding process. Very limited information on the buildings is provided by the authors, and the study was primarily effective for illustrating the manner in which buildings in the stock consumed electricity.

2.3 Classification of smart meter data for building characterization

As mentioned previously, works in smart meter data analytics for the purpose of building characterization are limited, primarily due to the lack of appropriate smart meter data sets for model training. Specifically, large residential electricity metered data sets accompanied by building parameters such as the surface area, building type, or the number of occupants, are rare.

Recently a large open-source data set of smart meter data with building meta-data has been made available for 1636 non-residential buildings in the U.S. (Miller et al.

2020). Najafi et al. (2021) performed a feature analysis study on the data set using the Random Forests classification algorithm to predict principal building use, performance and operations strategy. The results show the importance of feature selection and the possible classification accuracy that can be obtained by varying the features, but are restricted to non-residential buildings.

The Irish Social Science Data Archive (ISSDA) Commission for Energy Regulation (CER) data set of Irish residential dwellings with over 4000 homes with smart meter data contains mainly demographic data, though building surface area, number of occupants and building type were also included (CER 2012). Beckel et al. (2014) performed a classification study on the CER data set, which showed that linear discriminant analysis could predict the various class categories with classification accuracy between 35% and 80%. The large limitation of the study was that the heating and cooling electricity consumption were not included in the data, as these loads were covered by other energy sources.

Carroll et al. (2018) performed a study using the Neural Networks and Elastic Net Logistics machine learning techniques on smart meter data, again for the purpose of household demographic classification in Ireland. The CER data set was used similarly to Beckel et al. (2014), though with different machine learning techniques. What was particularly interesting about the Carroll et al. (2018) study was how they tested a combination of 21 different feature values representing different aggregated electricity consumption values. While some households with a lower number of occupants could be accurately classified, in general it was difficult to identify the appropriate demographic class category with a high degree of accuracy, at least for the techniques studied.

What the literature reveals is that there are few examples of classification on electricity smart meter data that evaluate the capability to predict a building class category, such as the heated surface area, based only on the electricity consumption values. While many machine learning classification algorithms exist, this study focuses on evaluating linear discriminant analysis (LDA) for a wide variety of cases, to test the robustness and the possible effects of using LDA on very large data sets.

2.4 Linear discriminant analysis

Linear discriminant analysis (LDA) is a robust classification algorithm that uses linear projection to establish a decision boundary between two or more data groups. This is accomplished by choosing a projection line perpendicular to the decision boundary that best separates the data groups by maximizing the distance between the means of the data groupings and minimizing the variance of the data sets. A discriminant function $\delta_c(x)$ can be determined for a set of data for a class of category c using Equation (1), where the goal is to determine the maximum value of $\delta_c(x)$.

$$\delta_c(x) = \boldsymbol{x}^T \boldsymbol{K}^{-1} \boldsymbol{\mu}_c - \frac{1}{2} \boldsymbol{\mu}_c^T \boldsymbol{K}^{-1} \boldsymbol{\mu}_c + log(p_c) \tag{1}$$

where $\delta_c(x)$ is the discriminant function, $\boldsymbol{x}$ contains the classification data, $\boldsymbol{\mu_c}$ are the mean values of the data set, $\boldsymbol{K}^{-1}$ is the inverse of the pooled covariance matrix, and p_c is the probability of a new data point belonging to class category c. A more detailed description and derivation of the components of Equation (1) is provided in Appendix 1.

Equation (1) expresses the projection of the mean and covariance of the data sets on a projection axis and establishes a decision line by maximizing the term to the right of the equal sign. Equating the linear projection equations for two class categories ($\delta_{c1}(x) = \delta_{c2}(x)$) results in a linear decision boundary that can be used to classify new data points as either class category c_1 or c_2, for example. An example of LDA applied

to energy consumption data with the derivation of the key equations is provided in Appendix 1 for interested readers. The example serves to illustrate how LDA can be applied to energy data in the context of predicting information about buildings.

2.5 Summary and paper organization

As illustrated in the literature review, machine learning has been applied to address a variety of building energy problems. Supervised machine learning classification on electricity smart meter data has been the subject of very few studies, primarily due to the fact that no appropriate data sets exist for predictive model training of building characteristics using metered data. A new study on machine learning classification using linear discriminant analysis on a virtual smart meter data set is therefore presented.

This paper is organized in a number of sections, with the goal to illustrate the supervised machine learning classification of electricity smart meter data. First, the data set used for predictive model training and development is described. Next, the classification results are presented. A general discussion is then provided to elaborate on the outcome of the study and compare the results to previous studies. Finally, some concluding remarks are provided.

3 Data set description

As described in the literature review, residential data sets for predictive model development are very limited in scope. A virtual smart meter (VSM) data set for residential buildings was developed by Neale et al. (2020a), which consists of 200,000 homes with a variety of known physical characteristics. The data set is available for anyone to download (Neale, Kummert, and Bernier 2020b). The VSM data allows for a variety of parameters and house types to be evaluated.

The virtual buildings and VSM data were generated using building energy simulations with parameters determined based on probability distributions using available building stock details. The 200,000 houses represent a subset that is relatively close to the distribution of approximately 2 million single-family homes across the province of Québec, Canada. Each home is represented by electricity consumption values for a full year at 15-minute intervals, i.e. 35,040 data points per building, and 21 physical characteristics, such as building location, heated surface area, building envelope thermal resistance and more. The building parameter classes and categories are described in Table 1 (Neale, Kummert, and Bernier 2020a).

Table 1. VSM data class category descriptions (adapted from Neale et al. 2020). PMF: probability mass function, UPD: uniform probability distribution (i.e. no prior knowledge for the building stock). Values in square brackets represent the median value for that category.

Class	**Categories**		**Type of distribution**	**Description**
	#	**Values**		
Location	7	1: Rimouski 2: Saguenay 3: Québec city 4: Sherbrooke 5: Trois-Rivières 6: Montréal 7: Gatineau	PMF	Region in the province of Québec, Canada, where the building is located.
Building dimensions and orientation				
Building type	4	1: Single-detached home (DET) 2: Row house (ROW) 3: Semi-detached home (SDH) 4: Other single-attached (OSA)	PMF	Type of home.
Aspect ratio	5	1: 0.8 2: 0.9 3: 1.0 4: 1.1 5: 1.2	UPD	Aspect ratio of the home, which refers to the ratio of the width (street-facing dimension) to the length.
Surface area (m^2)	5	1: 56-93 [75] 2: 93-139 [115] 3: 139-186 [160] 4: 186-232 [210] 5: >232 [250]	PMF	Heated surface area bins, from smallest to largest. Note that in addition to the surface area category, the exact surface area of the house within that bracket is also provided in the VSM data set. Values in square brackets represent the mean surface area for that bin.
Window-to-wall ratio	3	1: 0.1 2: 0.15 3: 0.20	UPD	Ratio of window surface area to wall surface area.

Class	Categories		Type of distribution	Description
	#	Values		
Building rotation	4	1: 0° 2: 90° 3: 180° 4: 270°	UPD	Rotation of the building with respect to south (90° increments).
Building adjacency	4	1: No adjacent buildings 2: Eastern wall adjacent 3: Western wall adjacent 4: Both eastern and western walls adjacent	UPD	Configuration of outdoor walls directly adjacent to another building. "Eastern" and "Western" are in reference to the front of the home being south facing.
Floors	2	1: 1-storey 2: 2-storey	PMF	Number of floors in the home.
Building envelope				
Wall thermal resistance (m^2KW^{-1})	4	1: 0.5-1.5 [1.0] 2: 1.5-2.5 [2.0] 3: 2.5- 4.5 [3.0] 4: >4.5 [5.0]	PMF	Wall thermal resistance value. Values in square brackets represent the mean value for that bin.
Roof thermal resistance (m^2KW^{-1})	6	1: 0.5-1.5 [1.0] 2: 1.5-2.5 [2.0] 3: 2.5-3.5 [3.0] 4: 3.5-4.5 [4.0] 5: 4.5-5.5 [5.0] 6: >5.5 [8.0]	PMF	Roof thermal resistance value. Values in square brackets represent the mean value for that bin.
Foundation thermal resistance (m^2KW^{-1})	4	1: 0.5-1.5 [1.0] 2: 1.5-2.5 [2.0] 3: 2.5-3.5 [3.0] 4: 3.5-4.5 [4.0]	PMF	Foundation thermal resistance value. Values in square brackets represent the mean value for that bin.
Overall building thermal resistance (m^2KW^{-1})	3	1: <1.56 2: 1.56-2.25 3: >2.25	UPD	Derived from the roof, wall and foundation thermal resistance values.

Class	Categories		Type of distribution	Description
	#	Values		
Air leakage area (cm^2 @4Pa)	5	1: 248 2: 406 3: 556 4: 775 5: 1426	PMF	Used to characterize air infiltration.
Window glazings	3	1: Single-glazed windows 2: Double-glazed windows 3: Triple-glazed windows	PMF	Number of window glazings used in the building.
Heating, air conditioning and domestic hot water				
Air conditioning	3	1: No air conditioning 2: Air-source heat pump 3: Window air conditioner	PMF	Air conditioning system used in the building, if any.
Heat pump	2	1: No heat pump 2: Air-source heat pump + Auxiliary	PMF	Heat pump type.
Auxiliary heating type	2	1: Electric 2: Non-electric	PMF	Auxiliary heating system type.
Occupancy information				
Occupancy profile number	15	1-15: Occupant profiles numbered 1 to 15	UPD	Stochastic occupancy load profiles used in the building simulation. 15 distinct stochastic profiles were used.
Number of occupants	5	1: 1 occupant 2: 2 occupants 3: 3 occupants 4: 4 occupants 5: 5 occupants	PMF	Number of occupants in the home.
Other parameters				
DHW type	2	1: Electric 2: Non-electric	PMF	Domestic hot water heater type.
Pool	2	1: Pool 2: No pool	PMF	Pool type.
Spa	2	1: Spa 2: No spa	PMF	Spa type.

Each house from the Neale *et al.* (2020a) VSM data set is accompanied by the corresponding category for the input parameters described in Table 1. The house's electricity consumption is therefore paired with a number of different building parameters that can be used for predictive model development.

3.1 Example data

The electricity consumption of a house depends on the combination of the parameters described in Table 1. In order to present classification results, it is relevant to discuss the smart meter data used to train the predictive models. If the smart meter data from the VSM data set is aggregated to annual electricity consumption it can be plotted in terms of the heated surface area category, as illustrated in Figure 2.

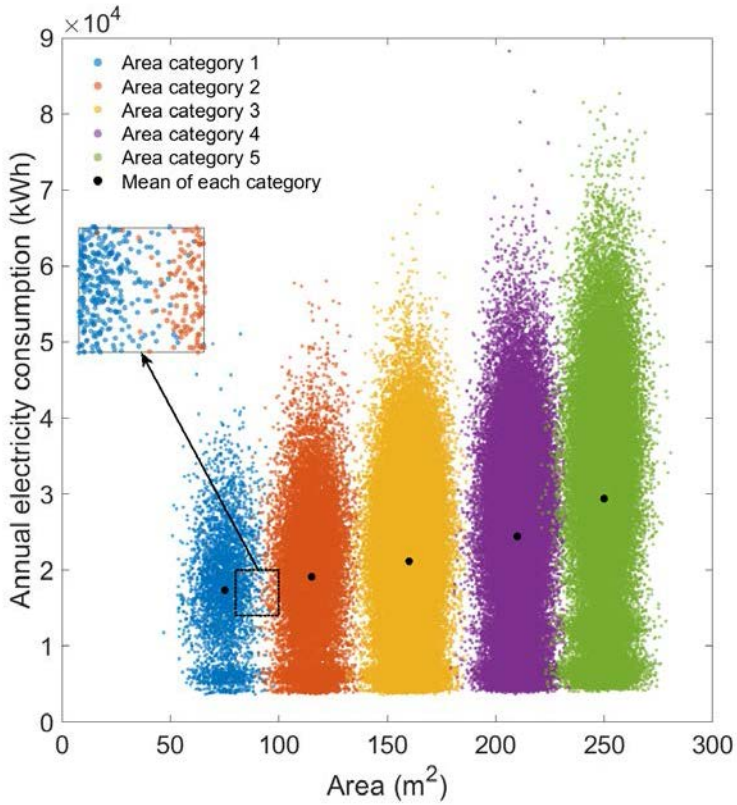

Figure 2. Annual electricity consumption of VSM data sorted by surface area category. Each point represents one house with distinct characteristics.

The data in Figure 2 illustrates how each heated surface area category has a wide range of annual electricity consumption values from the data set. There are gaps

between the surface area categories because the range of values for each category were generated with a Gaussian distribution, making fringe values for a given category less likely, as illustrated in the zoomed-in portion of Figure 2. While the categories are somewhat distinct in the figure, it is difficult to determine the size of a house with only the annual electricity consumption. For example, there are houses in all size categories with electricity consumption equal to 20,000 kWh, which indicates that further information is required to accurately classify the houses based on annual electricity consumption alone.

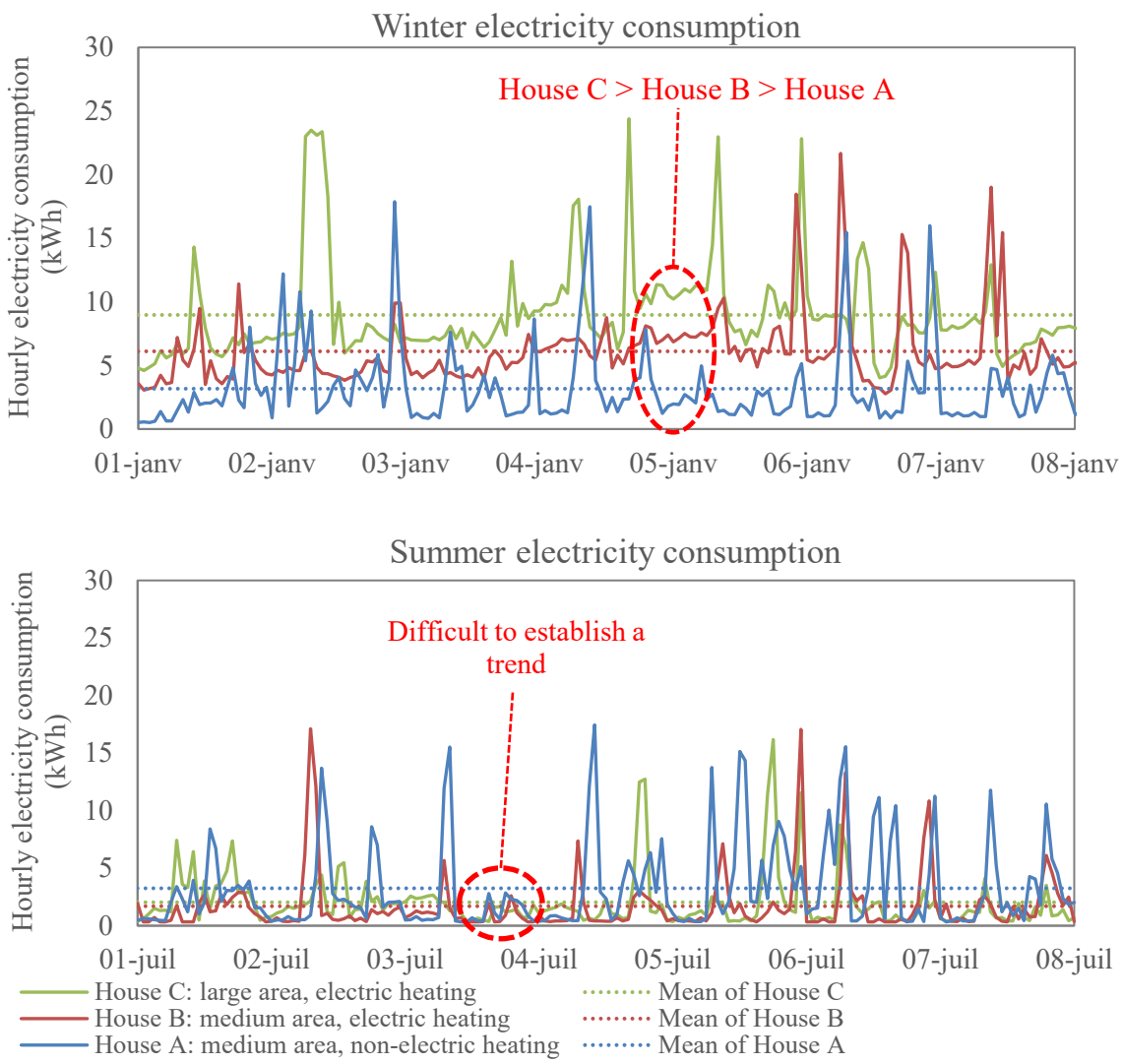

Figure 3. Electricity consumption for houses with different characteristics for January (top) and July (bottom) hourly data.

The electricity consumption values at hourly aggregation are illustrated in Figure 3 for three houses from the VSM data set for a week of data in January and July. A medium house (surface area category 3) without electric heating is shown, which provides an understanding on an electricity consumption profile based primarily on internal loads. Second, a medium house with electric baseboard heating illustrates the variation in electricity consumption when the heating load is included. Finally, a large house (surface area category 5) is compared to the others, with comparatively higher electricity consumption due to the increased overall loads.

The hourly-aggregated data in Figure 3 shows many peaks and valleys as the internal loads and the outdoor conditions vary. As indicated in the figure, distinct differences are noticeable in the winter electricity consumption based on the characteristics of the houses, such that visually they can be distinguished based on their size (medium vs. large) or by their heating system type (electric vs. non-electric). These differences are what the classification process aims to identify and associate with the various class categories. The summer electricity consumption illustrates that the profiles for the three illustrated houses are similar and not distinguishable by size or by heating system type, which is logical given the lack of heating load in July for the given building stock.

4 Classification methodology

As the data in Figure 2 illustrates, annual electricity consumption is insufficient for visual classification of homes based on the surface area category, except for extreme cases. A predictive model would need to distinguish the impact of the other characteristics by looking at the electricity smart meter data without knowing those other class categories. For example, a house with 5 occupants has higher variable internal loads than a house with 1 occupant, which should aid in distinguishing between

a small house with a large family (high base load, lower heating load) when compared to a large house with a single occupant (low base load, higher heating load).

Linear discriminant analysis as a classification technique for smart meter data is evaluated using a set of 200,000 virtual buildings with a variety of known geometries, internal loads, and heating, ventilation and air-conditioning (HVAC) system parameters. First, a brief description of the methodology behind the predictive model development is provided. The classification accuracy results for a number of scenarios are presented for linear discriminant analysis. Some specific cases are illustrated in more detail. The time required to produce predictive models based on the number of features is discussed. The impact of the number of buildings is then presented, which illustrates whether 200,000 buildings are required to accurately classify building parameters for the studied residential building stock. The effectiveness of the predictive models developed with the VSM data set are evaluated by applying real smart meter data to the models and comparing the predicted building parameters with known values. Finally, electricity consumption data for a small set of real houses are used to evaluate the prediction capability of the developed models.

4.1 Predictive model development methodology

Linear discriminant analysis is evaluated as a method to perform machine learning classification of smart meter data using building characteristics as response variables for electricity consumption predictors. A predictive model is developed using the following general methodology:

1. Select the number of buildings, i.e. 200,000 dwellings.
2. Select the number of features, i.e. hourly annual data has 8760 values and therefore that many features.
3. Build the predictor data matrix from the feature data of each building.

4. Build the response vector from the class categories of each building for the studied class, i.e. the surface area bin for the Area class for each dwelling.
5. Develop the predictive model from the predictor and response data.

The process above is repeated for a variety of configurations. The Matlab Statistics and Machine Learning Toolbox is used for all classification results in this study (Mathworks Inc. 2018).

4.2 Classification accuracy

The accuracy of a predictive model is determined based on the number of correct predictions using a validation scheme, as described in Equation (A-*12*). Predictive models are developed using linear discriminant analysis (LDA) for each building parameter included in the VSM data set. In addition, multiple feature scenarios are presented to study the effect of aggregating the VSM data on the classification accuracy. Scenarios I through IV reflect data for a full year at different time aggregation intervals (monthly, weekly, daily, hourly). Classification for a full year of smart meter data with subhourly values was not found to be possible due to computer memory limitations and the size of the resulting matrix equation to resolve the classification problem. Scenarios V through VIII are classification results for the month of January with different features (weekly, daily, hourly, and subhourly). Scenarios IX through XII reflect the same feature combinations, but for the month of July. The months of January and July were chosen to evaluate the impact of reducing the data set size and to test the prediction capability of the classification algorithms for building parameters with low or no impact of those parameters during those periods. For example, evaluating the classification of air conditioning using winter data should result in poor classification, since single-family homes typically have zero air conditioning load during the winter months. In addition, it was possible to use subhourly data for the monthly cases, since

the number of features was significantly smaller (2976 features) than the annual case (35,040 features). Note that all 200,000 buildings were used for classification, though the impact of the building set size is presented further in the results.

The classification accuracy (CA) results are presented in Table 2. Each accuracy value, from 0 to 1, reflects the prediction accuracy for a single predictive model. A value of 0.9 indicates that the class category of 90% of buildings in the data set were correctly predicted with that predictive model. The color scale in the table reflects the range of values within that scale, with red representing close to 0, and green representing close to 1. The best result for each class is indicated with a black border, favoring cases with less features if there is a tie.

The CA results are compared to the accuracy of performing a random guess (RG), which is based on the chance of guessing correctly without knowing any details about the building stock parameters. The RG value is simply $1/n_{cat}$, where n_{cat} is the number of categories in that class, which should be the absolute minimum threshold for classification accuracy. The CA results are above the RG in all cases, indicating the classification algorithm is better than blind guessing. Since the probability distributions for the VSM data set are provided by Neale et al. (2020a), the value of a random guess based on prior knowledge (RG_{PK}) can also be determined based on those distributions. This value accounts for the probability of some class categories being more prevalent, which results in a higher chance to guess the correct outcome. Not all parameters had prior knowledge when the VSM data set was developed and therefore some RG_{PK} values are not included in Table 2.

In order to facilitate the comprehension of the results in Table 2 an example is provided. The *Area* class has five categories, which represent five surface area bins described in Table 1. The probability of randomly guessing an area bin without any

prior knowledge would be 1 in 5, or RG = 0.20. The *Area* class values were generated using a probability mass function that depended on the type of home (detached, semi-detached, etc.) and the number of floors in the home (Neale, Kummert, and Bernier 2020a). Since those profiles are available, the probability of correctly guessing the class category can be calculated, and in the case of the *Area* class $RG_{PK} = 0.279$, slightly better than the blind guess value. The classification accuracy for Scenario I – monthly data for a full year of electricity consumption – is equal to 0.457, which is somewhat better than randomly guessing the category. By increasing the number of features to use hourly data, which corresponds to Scenario IV in Table 2, the accuracy improves to 0.793. For this case, the predictive model correctly predicts the surface area category for 4 out of 5 homes in the data set.

To expand upon the *Area* class example, the confusion matrices (CM) for Scenarios I and IV are illustrated in Figure 4. The CM provides an understanding of the proportion of correct and incorrect predictions of the predictive model for each category. If a model results in only one category being correctly classified then the model is not very useful when the goal is to identify building characteristics spanning multiple categories.

True class category	1	2	3	4	5	TP	TN
1	**0**	0	3827	5	98	**0.000**	1.000
2	0	**0**	18286	148	1664	**0.000**	1.000
3	0	0	**63248**	2731	8716	**0.847**	0.153
4	0	0	35625	**4620**	12994	**0.087**	0.913
5	0	0	21109	3442	**23487**	**0.489**	0.511

Predicted category

True class category	1	2	3	4	5	TP	TN
1	**2478**	755	654	43	0	**0.631**	0.369
2	887	**13194**	5646	371	0	**0.656**	0.344
3	32	1496	**69505**	3654	8	**0.931**	0.069
4	1	0	4971	**43859**	4408	**0.824**	0.176
5	0	0	1674	10632	**35732**	**0.744**	0.256

Predicted category

Figure 4. Confusion matrix for Scenario I (left) and Scenario IV for the Area class categories, labelled 1 through 5. TP: true positive, TN: true negative. Bolded values illustrate the correctly predicted cases.

Correct predictions in a confusion matrix are placed in the main diagonal where *Predicted category = True category*. The confusion matrix for Scenario I in Figure 4 illustrates that categories 1 and 2 were incorrectly assigned to categories 3 to 5. This indicates that with monthly features, the difference in electricity consumption is too subtle for the predictive model to distinguish the smallest house size categories. The True Positive (TP) and True Negative (TN) columns indicate the proportion of correctly and incorrectly predicted buildings in the data set, respectively. In Scenario IV there is a much more even spread of building predictions, and in most cases the results are within one size category of being correctly predicted. The corresponding classification accuracy values can be calculated from the confusion matrices by summing the diagonal values, which are the correct predictions, and dividing by the number of houses in the data set, which in this case is equal to 200,000. For example, the case on the left of Figure 4 has 91,355 correct predictions (63,428 + 4620 + 23,487) out of 200,000 total houses, and therefore 0.457 classification accuracy.

Table 2. Classification accuracy results. n_{cat}: number of category values for that class, RG: random guess, RG_{PK}: random guess with prior knowledge. Results with a dark outline indicate the best result for that class.

Scenario:	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**			
Period:	**1 year**	**1 year**	**1 year**	**1 year**	**Jan.**	**Jan.**	**Jan.**	**Jan.**	**July**	**July**	**July**	**July**			
Interval:	**Monthly**	**Weekly**	**Daily**	**Hourly**	**Weekly**	**Daily**	**Hourly**	**15-min.**	**Weekly**	**Daily**	**Hourly**	**15-min.**			
Features:	**12**	**52**	**365**	**8760**	**4**	**31**	**744**	**2976**	**4**	**31**	**744**	**2976**	n_{cat}	**RG**	RG_{PK}
Location	0.840	0.938	0.970	0.973	0.727	0.929	0.964	0.965	0.598	0.717	0.948	0.956	7	0.143	0.386
Physical properties															
Building type	0.801	0.802	0.859	0.928	0.801	0.799	0.861	0.899	0.801	0.801	0.807	0.812	4	0.250	0.662
Aspect ratio	0.201	0.201	0.201	0.206	0.201	0.201	0.202	0.200	0.201	0.201	0.199	0.200	5	0.200	*
Area	0.457	0.459	0.574	0.793	0.461	0.456	0.649	0.775	0.375	0.416	0.540	0.577	5	0.200	0.279
WWR	0.464	0.516	0.589	0.829	0.368	0.469	0.750	0.782	0.348	0.364	0.525	0.572	3	0.333	*
Rotation	0.259	0.265	0.281	0.286	0.253	0.259	0.290	0.286	0.249	0.250	0.268	0.270	4	0.250	*
Adjacent buildings	0.801	0.802	0.843	0.894	0.801	0.799	0.845	0.872	0.801	0.801	0.807	0.812	4	0.250	0.655
Number of floors	0.682	0.720	0.877	0.940	0.635	0.689	0.907	0.922	0.566	0.606	0.808	0.843	2	0.500	0.510
Building envelope															
Wall thermal resistance	0.590	0.607	0.667	0.838	0.590	0.591	0.755	0.792	0.590	0.590	0.634	0.657	4	0.250	0.432
Roof thermal resistance	0.241	0.266	0.360	0.656	0.238	0.241	0.488	0.552	0.215	0.236	0.301	0.334	6	0.167	0.186
Foundation thermal resistance	0.628	0.698	0.849	0.908	0.538	0.600	0.841	0.853	0.482	0.544	0.803	0.819	4	0.250	0.420
Overall thermal resistance	0.491	0.539	0.632	0.756	0.429	0.464	0.677	0.711	0.363	0.438	0.622	0.643	3	0.333	*
Leakage	0.351	0.418	0.590	0.655	0.332	0.381	0.658	0.642	0.217	0.248	0.436	0.482	5	0.200	*
Window glazings	0.902	0.900	0.897	0.959	0.902	0.902	0.932	0.949	0.902	0.902	0.895	0.900	3	0.333	0.818
HVAC															
Air conditioning	0.777	0.817	0.902	0.981	0.739	0.772	0.819	0.817	0.698	0.763	0.926	0.953	3	0.333	0.509
Heat pump	0.952	0.948	0.988	1.000	0.910	0.946	0.997	0.999	0.838	0.843	0.923	0.924	2	0.500	0.729
Auxiliary heating	0.986	0.993	1.000	1.000	0.987	0.984	1.000	1.000	0.731	0.799	0.959	0.959	2	0.500	0.607
Occupancy details															
Occupants	0.506	0.799	1.000	1.000	0.370	0.607	1.000	1.000	0.416	0.748	1.000	1.000	5	0.200	0.246
Profile number	0.270	0.892	1.000	1.000	0.139	0.622	1.000	1.000	0.252	0.808	1.000	1.000	15	0.067	*
Other parameters															
DHW type	0.960	0.963	1.000	1.000	0.957	0.958	1.000	1.000	0.772	0.817	1.000	1.000	2	0.500	0.649
Pool	1.000	1.000	1.000	1.000	0.835	0.835	0.835	0.833	0.992	0.994	1.000	1.000	2	0.500	0.725
Spa	0.919	0.950	1.000	1.000	0.900	0.900	1.000	1.000	0.903	0.904	0.910	0.894	2	0.500	0.820

* No prior knowledge for the probability distribution for the category values existed for this parameter, therefore no improvement can be made over simply randomly guessing the outcome.

The results in Table 2 illustrate a wide range of classification accuracy values that depend on the scenario and the class. Some classes are not well classified, such as the building rotation or aspect ratio. LDA does not perform any better than randomly guessing for these cases, which indicates that these parameters in the VSM data set have little influence on the smart meter data. Other classes have significantly better accuracy, especially as the number of features increases. The number of occupants and the occupancy activity level is easily classified with daily and hourly data, which is likely due to the programmed nature of the profiles and is one of the limitations of simulation-based occupancy models.

In order to visualize the impact of the number of features on the classification accuracy, the data in Table 2 can be expressed in graphical form. The CA results for the location, area, air infiltration rate and overall envelope thermal resistance are illustrated in Figure 5.

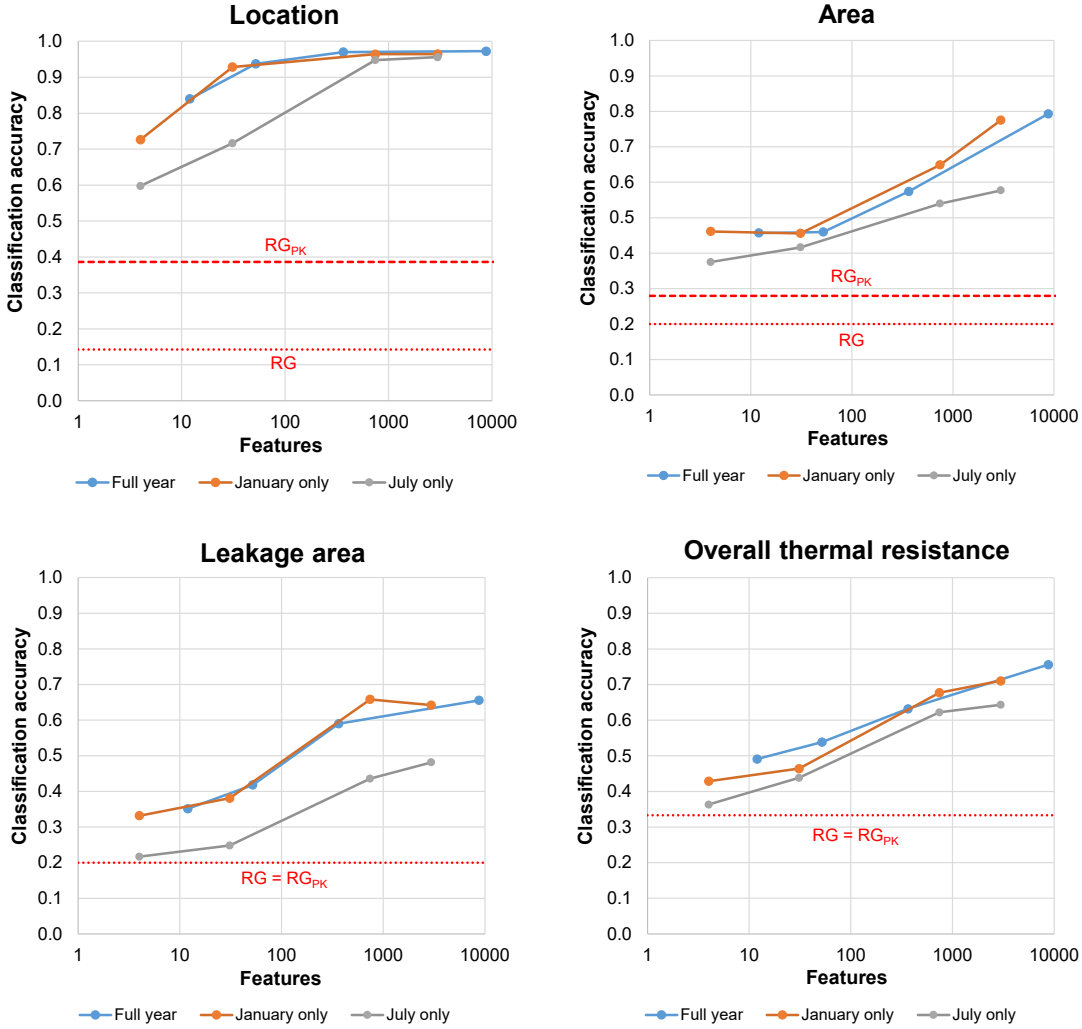

Figure 5. Classification accuracy per feature for the location, heated surface area, air infiltration and overall thermal resistance parameters. RG: random guess, RGPK: random guess based on prior knowledge.

The accuracy for LDA predictive models typically increases with a higher number of features, indicating in many cases additional granularity in the electricity consumption is beneficial for classification for building parameters. In some cases, the accuracy reaches a plateau at higher feature values, indicating that there is little gain for increasing the complexity of the predictive model, such as for the Location parameter. For other parameters, such as the Area, increasing the data granularity further could be beneficial, but at significant computational cost.

Finally, the classification accuracy for January and July smart meter data is illustrated in Figure 6 for all cases where $abs(CA_{jan} - CA_{july}) > 0.005$, i.e. if there is

a meaningful difference between the January and July classification results with the same number of features. Other cases where $-0.005 \leq (CA_{jan} - CA_{july}) \leq 0.005$ are set equal to ±0.005 for visibility but are considered to have a negligible difference for practical purposes. Figure 6 illustrates whether there is a difference between summer and winter classification by qualitatively comparing the class results. Each bar in Figure 6 represents one pair of CA results that demonstrate the improvement in accuracy using summer or winter data for classification. Results on the negative x-axis illustrate cases where the July data resulted in better classification, while results on the positive x-axis demonstrate cases where January data offered better outcomes. The bar length is a relative difference and thus does not illustrate the absolute accuracy, though these can be obtained in Table 2.

The results in Figure 6 demonstrate a logical link between the class and the preferred data set to use for classification, when choosing between available winter or summer data. Parameters that impact the heat gain and losses in a home are better classified using January data, as the larger temperature difference between indoor and outdoor leads to proportionately larger heat losses, and therefore more easily detectable differences between the class categories. Similarly, systems related to the heating load of a house are better classified with winter data. Domestic hot water loads, which are tied to ground water temperature, are also better classified in winter.

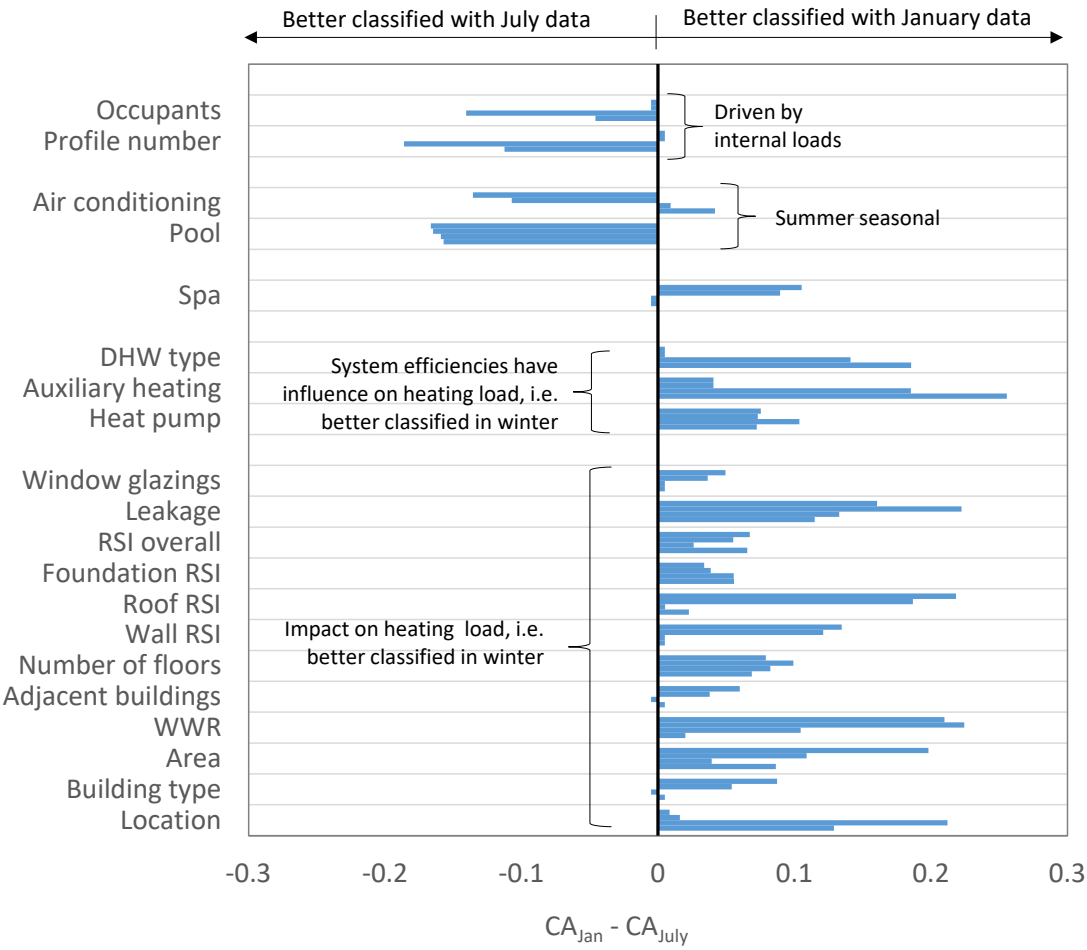

Figure 6. Comparison of January and July classification accuracy

Some classes are better classified with summer data, such as the seasonal air conditioning and pool loads. Occupancy-driven internal loads are easier to detect in summer periods due to the lower or non-existent effect of the cooling loads during these periods. The classification process appears to have an easier task at differentiating different occupancy patterns and number of residents using July data. The results in Figure 6 illustrate that it is worth considering what parameter is being classified when choosing between seasonal smart meter data sets.

4.3 LDA predictive model development time

Increasing the number of features in a classification problem will exponentially increase the size of the matrix equation required to solve for the classification boundaries, resulting in increased computation time. The time to compute each predictive model

result in Table 2 for each feature scenario was recorded. The average time per feature across all classes is illustrated in Figure 7. Classes with a higher number of categories tend to take longer as there are additional classification boundaries between each category. All predictive model development was performed on an Intel Core i9-7920X processor @2.9 GHz, 128 GB of RAM @2133 MHz and a SATA III solid-state hard drive.

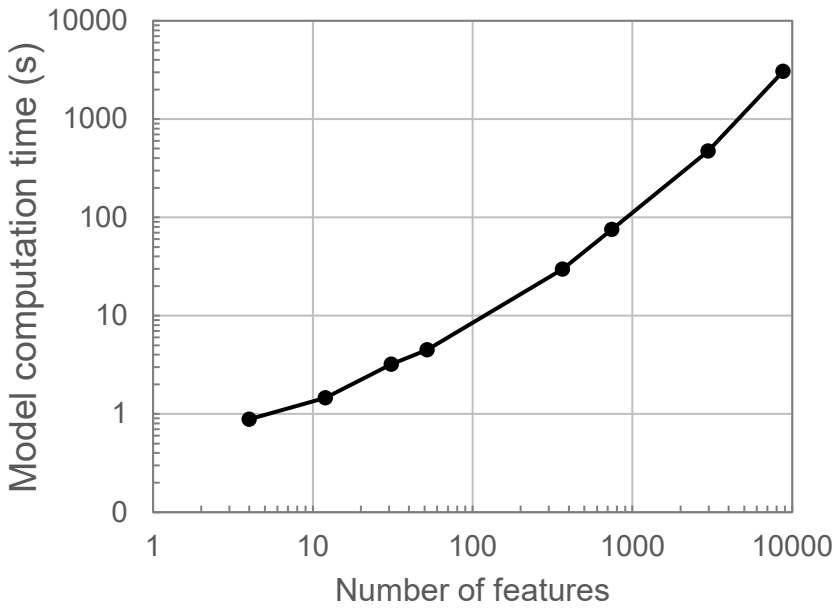

Figure 7. Average model computation time based on the number of features

As the number of features increases, the time to compute the predictive model increases exponentially. For an annual predictive model with hourly data, the average computation time is approximately 50 minutes. It should be noted that using 5-fold cross-validation significantly increases the overall time, as it repeats the predictive model process for each fold, using 80% of the data for training and 20% for validation. In addition, the parallel processing features of the Matlab Parallel Computing Toolbox are used to expedite the calculation process for the presented results (Mathworks Inc. 2018). The chief limitation of the predictive model development is not the time to

resolve the model but the quantity of random access memory (RAM) required to store and process the data, with 128 GB RAM being insufficient for some cases.

4.4 Impact of data set size on classification accuracy

The ratio of the data set size to the number of features influences the classification accuracy (Hua et al. 2005). In the case of the present study, the number of buildings in the VSM data set determines the size and the aggregation interval determines the number of features. By testing various subsets of buildings for monthly, weekly, daily and hourly electricity consumption, the range of classification accuracy values is illustrated in Figure 8. The y-axis describes the classification accuracy and the x-axis, which is on a logarithmic scale, describes the number of buildings used to develop the predictive models for classification. The range of values represented by the shaded area illustrates the effect of developing a predictive model with different sets of buildings. The smaller the amount of buildings, the more likely the chance of a statistically unrepresentative sample, which results in highly variable classification accuracy.

For example, Figure 8(a) shows that testing various sets of 10 buildings using monthly data for the Area class resulted in classification accuracy values ranging from 0.00 to 0.78 (i.e. the range of values plotted on the y-axis). This is due to the effect of the very small sample of buildings and high variability in characteristics of those buildings. Conversely, sets of 5000 buildings with daily data (notation in Figure 8(c)) resulted in a much smaller range of values for the Area class, from 0.486 to 0.516. The values at the extreme right of each curve in Figure 8 correspond to those in Table 2, for Scenarios I through IV for the Area class. For the latter example of Figure 8(c), which corresponds to Scenario III (1 year of daily features) for Area, the classification accuracy is equal to 0.574.

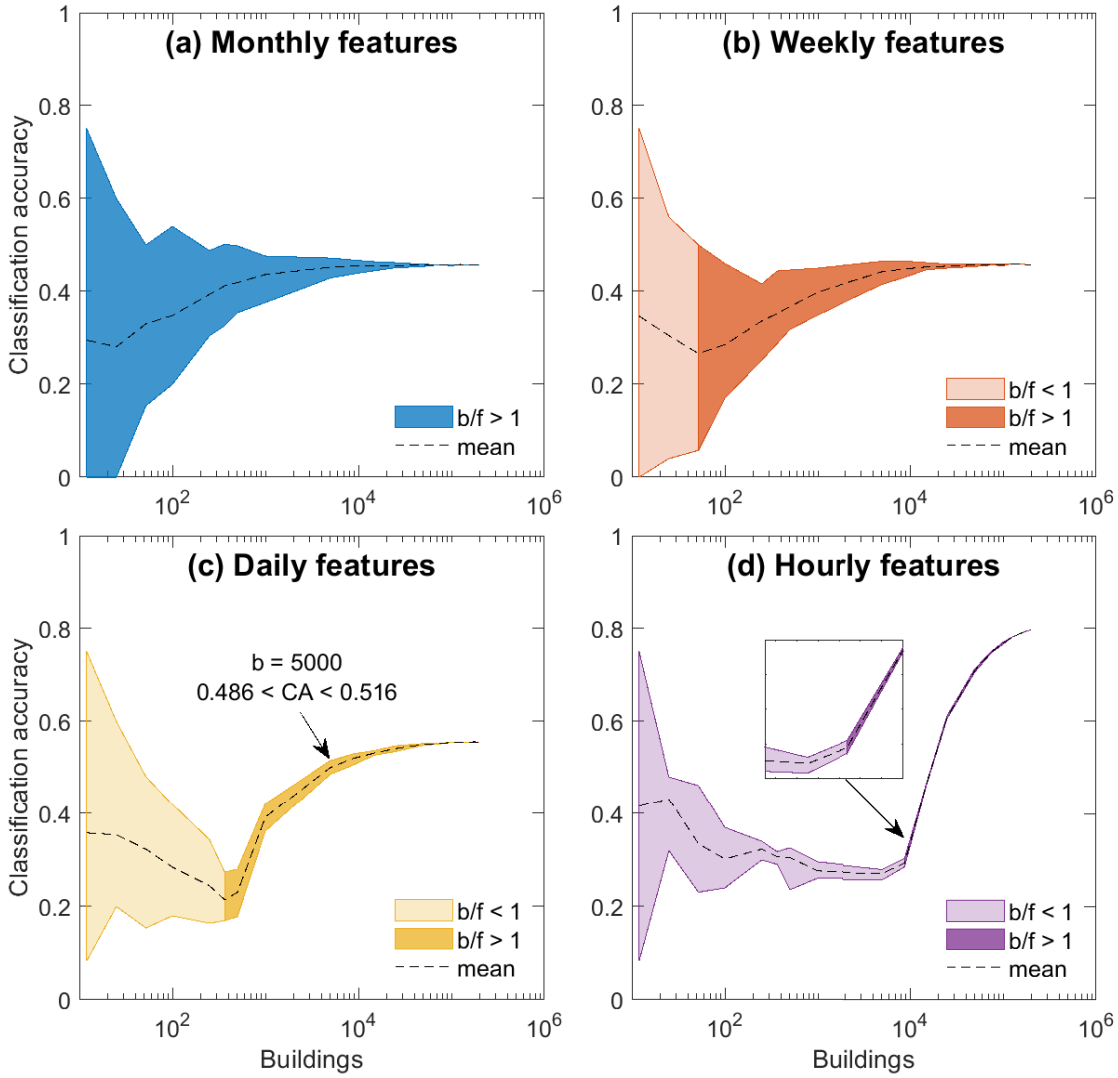

Figure 8. Area classification accuracy by building data set size for (a) monthly (12), (b) weekly (52), (c) daily (365) and (d) hourly (8760) features. b: number of buildings in data set, f: number of features, CA: classification accuracy.

The large range in classification accuracies for a smaller number of buildings can be explained by the likelihood of obtaining the correct predictions by chance. As the number of buildings increases and the characteristics diversify in the data set, the predictive model development stabilizes. This transition occurs when the number of buildings (b) is approximately equal to the number of features (f), or at $b/f \cong 1$, which is represented in the figure as the transition point between the lighter and darker shaded areas in each graph. The monthly case has only one shaded region as the classification process requires more than 12 buildings to be effective, and therefore $b/f > 1$ for all developed predictive models for this case. As the number of buildings used in the data

set further increases, the range of classification accuracy values narrows and the mean accuracy steadily increases, as depicted in Figure 8.

For smart meter data, it is therefore important to have a sufficiently sized sample of buildings to train the model if reliable classification results are desired. With hourly data, this indicates at least 8760 buildings with a variety of characteristics are required. Regardless of the data aggregation scheme used, the classification accuracy stabilized with additional buildings and increased for the daily and hourly feature cases. The curves illustrating the mean values of the accuracy demonstrate the increased performance based on the number of buildings.

4.5 Application of the developed predictive models on real smart meter data

The predictive model development process is illustrated in Figure 1, which describes how a developed model can be used to predict category values for new data. Using real smart meter data as inputs to a model developed with virtual data has some limitations. As an example, a virtual data set based on building energy simulations require occupancy models that inevitably differ from real occupants. These differences can result in unreliable classification, as the underlying assumptions in the virtual model can never perfectly match the reality. However, as mentioned in the literature review, there are currently no appropriate data sets based on real buildings to explore classification of building characteristics, and thus virtual smart meter data sets are the best option for now.

In order to present some of the limitations inherent to classification and guide future research in supervised machine learning of smart meter data, the predictive models developed and presented in Table 2 are used to predict building parameters for 30 houses with measured smart meter data, subsequently referred to as *real smart meter (RSM)* data. This general approach is described in Figure 1, where new predictor data is

input to a predictive model in order to determine the class category for that data. In this case, the RSM data is input to the developed models using the VSM data set, which provides an evaluation of the generalization of the predictive models (Shmueli 2010).

On average the RSM data was missing 3% of electricity consumption data, which were filled using recommended metered data processing techniques (Fowler et al. 2015). For some building characteristics, the true class values for the houses in the RSM data set are available as well, which are used to compare to the predicted categories. The data for the VSM and RSM sets were both for the calendar year 2016. The characteristics of the 30 RSM houses are presented in Table 3. All houses are located in *Location 5: Trois-Rivières* (Table 1).

Table 3. RSM house characteristics. AC: air conditioning. Cat: class category according to Table 1.

House	Area		Occupants	Air conditioning	Pool	Spa	Window glazings
	m^2	Cat					
1	191.4	4	4	No AC	Pool	No spa	Double
2	191.4	4	2	Heat pump	No pool	No spa	Double
3	198.2	4	4	No AC	Pool	No spa	Double
4	180.6	3	5	Heat pump	No pool	No spa	Double
5	106.7	2	2	No AC	No pool	No spa	Double
6	145.7	3	4	Heat pump	No pool	No spa	Double
7	162.8	3	1	No AC	No pool	Spa	Double
8	162.1	3	2	No AC	Pool	No spa	Double
9	214.2	4	5	No AC	Pool	Spa	Double
10	204.9	4	4	No AC	Pool	No spa	Double
11	229.5	4	5	Heat pump	No pool	Spa	Double
12	181.8	3	4	No AC	Pool	No spa	Double
13	334.5	5	6*	No AC	Pool	No spa	Double
14	258.7	5	4	No AC	No pool	No spa	Double
15	185.8	3	4	Heat pump	No pool	Spa	Double
16	139.4	3	2	Heat pump	No pool	Spa	Triple
17	204.0	4	5	No AC	Pool	No spa	Double
18	188.0	4	1	Heat pump	No pool	No spa	Double
19	152.2	3	2	No AC	No pool	Spa	Double
20	268.4	5	4	No AC	No pool	No spa	Double
21	179.8	3	2	Heat pump	Pool	No spa	Double
22	152.9	3	3	No AC	No pool	Spa	Double
23	151.1	3	3	Heat pump	No pool	Spa	Double
24	170.0	3	2	No AC	No pool	Spa	Double
25	188.8	4	3	Heat pump	Pool	No spa	Double
26	167.2	3	6*	Heat pump	Pool	No spa	Double
27	346.3	5	2	Heat pump	Pool	Spa	Double
28	330.3	5	4	No AC	No pool	No spa	Double
29	144.0	3	1	No AC	No pool	No spa	Double
30	188.1	4	4	Heat pump	No pool	No spa	Double

* The VSM data set only contained data for up to 5 occupants, therefore houses with 6 occupants are considered to have 5 instead.

The classification approach for the seven known RSM house parameters is described in Table 4. Class categories are assigned to each house based on the VSM parameters in Table 3. Classification accuracy is determined based on the similarity of the predicted category when compared with the true category, with exact matches described as "correct predictions" and with similar matches described as "close predictions". The definition of "close" varies by parameter and is included in the results to illustrate when classification obtains outcomes equal to or near the correct prediction. For example, if the number of occupants is predicted one higher or lower than a house's true occupancy, this is considered a close prediction. Some parameters, such as whether a pool is installed in the home, have no "close" option, as they are either correct or incorrect.

Table 4. House data set known parameters and definitions for a correct and close prediction

Parameter	Correct prediction	Close prediction
Location	Correct location predicted	Predicted as the correct location or another location with similar heating degree-days (HDD). ±130 HDD
Area	Correct area category predicted	Predicted as the correct area category or one size category larger or smaller. ±40 m^2
Occupants	Correct number of occupants predicted	Predicted as the correct number of occupants or one occupant more or less than the correct number. ±1 occupant
Air conditioning	Correct AC type predicted	Predicted as the correct air AC type, or if a heat pump predicted as a window air conditioner, or vice versa.
Pool	Presence of a pool correctly predicted	Not applicable
Spa	Presence of a spa correctly predicted	Not applicable
Window glazings	Correct number of window glazings predicted	Not applicable

The classification accuracy (CA) is determined as described in Equation (A-12), which is based on the correct predictions (CP) divided by the total predictions (TP). If applicable, close predictions are substituted in Equation (A-12) for the correct predictions. Classification accuracy for the RSM data is denoted as CA_{RSM}, which are

presented in Table 5. CA_{RSM} results are compared to the accuracy of randomly guessing the categories of each class based on the prior knowledge of the building stock from the VSM data set. The RG_{PK} is used as a reference since the RSM houses are part of the same building stock as the VSM data.

The results in Table 5 illustrate the classification accuracy when the real smart meter data is input to the predictive models developed with the VSM data and the predicted class category is compared to the real class category. As an example, an accuracy of 0.433 indicates that 13 out of 30 houses in the RSM data set had the category correctly predicted. This value can be directly compared to the RG_{PK} column to evaluate the performance of LDA when compared to a random prediction. If $CA_{RSM} > RG_{PK}$, the classification algorithm represents an improvement over a random guess.

Classification of the real smart meter data with linear discriminant analysis has variable accuracy depending on the class, the number of features, and the period used for the smart meter data. There is at least one scenario for each class that resulted in a better prediction than randomly guessing. The average CA improvement for the best scenario for each class is equal to 0.187 and ranges from 0.078 to 0.355. Scenarios with less features generally performed better, which indicates that aggregating the electricity consumption improves the classification accuracy. This is likely due to the way internal loads were generated in the VSM data set used to train the predictive models. Since it is unlikely to match occupant behavior to real data at a subhourly or hourly frequency, aggregating those data for classification seems to be the more reliable approach.

Table 5. Classification accuracy results for the real smart meter data set. Best classification results have bold text and borders.

Class	Prediction type	Scenario (*period-aggregation-features*)												RGPK
		1 year-M-12	1 year-W-52	1 year-D-365	1 year-H-8760	Jan-W-4	Jan-D-31	Jan-H-744	Jan-SH-2976	July-W-4	July-D-31	July-H-744	July-SH-2976	
Location	Correct	**0.433**	**0.433**	0.267	0.267	0.000	0.167	0.233	0.167	0.000	0.067	0.100	0.033	0.078
	Close	0.667	0.667	0.633	**0.700**	0.167	0.567	0.633	0.567	0.233	0.233	0.533	0.400	0.282
Area	Correct	0.467	0.233	0.167	0.167	**0.500**	**0.500**	0.133	0.100	0.467	0.400	0.133	0.133	0.300
	Close	0.833	0.533	0.433	0.500	**0.867**	**0.867**	0.367	0.433	0.833	0.767	0.400	0.433	0.433
Occupant	Correct	0.300	0.167	0.133	0.333	0.267	0.100	0.267	0.167	**0.400**	0.300	0.267	0.200	0.216
	Close	0.667	**0.700**	0.667	0.567	0.433	0.400	0.533	0.567	0.567	0.600	0.667	0.667	0.312
Air conditioning	Correct	0.333	0.300	0.500	0.533	0.500	**0.567**	0.367	0.333	0.533	0.533	0.500	0.433	0.488
	Close	0.433	0.433	0.500	**0.700**	0.500	0.567	0.400	0.467	0.533	0.533	0.533	0.467	0.033
Pool	Correct	0.600	0.600	0.600	0.600	0.600	0.600	0.500	0.600	**0.800**	0.667	0.567	0.467	0.563
Spa	Correct	0.500	0.400	0.533	0.500	0.667	0.633	0.400	**0.733**	0.600	0.533	0.467	0.467	0.650
Window glazings	Correct	**0.967**	0.900	0.600	0.100	**0.967**	**0.967**	0.267	0.167	**0.967**	**0.967**	0.100	0.100	0.873

In summary, linear discriminant analysis had mixed results predicting the class categories for a number of building characteristics for a real small data set. The combination of period and aggregation of the electricity consumption that resulted in the best classification result varied by parameter, which further supports the need for additional studies in classification of smart meter data. The data set of real houses used in the present study was quite limited in the number of houses available and the amount of known parameters that could be used for validation purposes. In addition, a non-negligible fraction of data was missing for the real houses, which certainly affects the classification prediction. The impact of the missing data is supported by the fact that aggregating the electricity data often resulted in better predictions. Nevertheless, LDA did demonstrate an improvement over random guessing for all parameters, at least for specific data scenarios. A larger, more detailed RSM data set would provide a better understanding of the link between the classification accuracy, number of features, data set size and number of buildings in the data set.

5 General discussion

The literature review illustrated the prevalence of machine learning in building applications and the lack of previous studies in classification of buildings based on smart meter data. The study of Beckel et al. (2014) compared multiple classification techniques to predict building characteristics using smart meter data, though most classes were related to occupancy. LDA predicted the floor area category (<100 m^2, 100 to 200 m^2 or >200 m^2) for homes in the Beckel study with an accuracy of 45%, compared with up to 80% in this study. For building type (detached or attached), Beckel's study classified houses with 60% accuracy, compared to 93% in this study. The number of occupants was predicted with approximately 70% accuracy, compared to 100% in this study. Carroll et al. (2018) performed a similar analysis as Beckel for

occupancy classification, averaging 61% accuracy with different classification algorithms.

This study improves upon previous classification works by systematically analysing the impact of a significant number of scenarios on the classification problem, which guides future classification modelers on the correct way of approaching smart meter data classification. The open-source VSM data set used to train the predictive models represents a new source of data for classification problems that has yet to be fully explored. Until a real smart meter data set with a variety of measured and surveyed building characteristics is released, the virtual data represents the best data set for smart meter classification studies.

While this paper provides a detailed evaluation of linear discriminant analysis using the VSM data, further work evaluating other classification algorithms using additional metrics would guide those seeking to perform supervised machine learning classification. Other algorithms may improve the results for smart meter data classification. When developing a virtual smart meter data set for classification, some parameters may not be worth attributing distinct class categories, such as differentiating between heat pumps and window air conditioners with similar coefficient of performance (COP) values. Care must be given when attributing class categories and when modeling specific physical behavior, such as the properties of windows installed in a house, as these can only be identified by classification if they were included in the original data set. In addition, a detailed monitoring campaign of real houses with surveyed building characteristics would greatly assist in the validation process of classification studies.

6 Conclusion

Building stock energy modeling requires a significant amount of information to accurately represent the wide range of building types. This paper seeks to illustrate how supervised machine learning classification with linear discriminant analysis (LDA) can accurately predict building parameters from electricity smart meter data. The virtual smart meter (VSM) data developed by Neale et al. (2020a) is a residential smart meter data set with detailed information on building characteristics, such as building surface area, thermal resistance of the building envelope, occupants, air leakage rate, etc. The VSM data was developed with classification in mind, and the present study uses LDA to evaluate the effectiveness of classification to predict building parameters based solely on electricity smart meter data. Data periods and aggregation intervals are varied to test a number of different feature combinations for each class.

Linear discriminant analysis can effectively classify electricity smart meter data, with classification accuracy values that depend on the parameter studied and the number of features. The building data set size has an important influence on the reliability of the classification outcome. At the very minimum, it is essential to have at least as many buildings (b) as the number of features (f) in the data set ($b/f > 1$). As this ratio increases, the classification accuracy for LDA tends to reach an asymptotic value when $b \gg f$. This indicates that for a building data set with highly variable characteristics and for parameters better classified with hourly or subhourly features, many buildings are required to develop a reliable predictive model.

Classification accuracy is related to the impact of a building parameter on the electricity consumption. Parameters such as building rotation and aspect ratio are not

well classified by LDA. This could be related to the way they are implemented in the building simulation environment used to create the VSM data set. Nevertheless, LDA performs no better than randomly guessing for these two particular parameters.

Other parameters had significantly better classification accuracy than randomly guessing and in some cases reached 100% accuracy, i.e. all 200,000 buildings had their class category accurately predicted by the predictive model. Classification accuracy is strongly tied to the number of features used to develop the model, with higher numbers of features generally resulting in higher accuracy. However, increasing the feature count significantly slows the predictive model development time and increases the memory requirements, as the equations required to resolve the classification problem scale exponentially. There is therefore a significant compromise between accuracy, computation time and computational resources.

One example of a parameter with high classification accuracy is the Occupants parameter. The VSM data set has 15 different profiles for 1 to 5 possible occupants, resulting in 75 different occupancy profiles. While this appears to be a high number of different cases, the classification algorithm can easily detect the differences between each number of occupants and between each profile. Practically speaking the profiles themselves are unlikely to correspond exactly to real house occupants, and so a developed predictive model based on occupancy simulations must be applied with caution.

Applying smart meter data from 30 houses to the developed predictive models resulted in variable classification accuracy. Classification was more effective for aggregate electricity consumption, leading to the conclusion that the stochastic loads of the virtual data set did not fully correspond with the real house occupants. The authors recommend using aggregated electricity consumption to more easily correspond

between modeled occupancy and real occupancy, should the need arise. A more detailed and more extensive real smart meter data set would allow for a better validation of the predictive models developed using the virtual set. Unfortunately, to the knowledge of the authors an appropriate data set for residential building classification does not exist, especially given the conclusions of this paper on the number of buildings required for reliable classification with higher numbers of features.

The results of this paper illustrate that classification has the potential to aid in the segmentation and characterisation of residential building stocks, provided a sufficiently detailed smart meter data set exists to train the models. This would directly benefit those seeking to develop building stock energy models but lack information about the buildings in the studied stock. Given the highly stochastic nature of residential electricity consumption, which depends on the individuals inhabiting the house, proper classification of real smart data using a virtual set of data is not guaranteed. It would be preferable to use a sufficiently large, detailed real smart meter data set with knowledge of the building characteristics to train the predictive model. The virtual classification results illustrates that building parameters can be predicted with a high level of accuracy with the electricity consumption only, which is a promising outcome as a source of data for future building stock energy modeling work.

7 Glossary

Features (f): *Features* are the number of data points representing a single building's energy consumption. By default a single building is represented by 35 040 features, which are electricity consumption data at 15 minute intervals for a full year. Other values are possible since the energy data can be aggregated, for example 365 features for daily-aggregated electricity consumption, or 12 features for monthly-aggregated data.

Predictors (p): The *predictors* represent the complete energy consumption data used to develop the predictive models. The predictor data set is a $[f \times b]$ matrix, where f is the number of features and b is the number of buildings, for example $[365 \times 1000]$ for a data set of daily energy use values of 1000 buildings.

Class (c) and category (cat): The *class* is the building parameter selected for classification, such as the building heated surface area or the location of the building. Each class is divided into a number of *categories*, which typically represent bins of values, or discrete values, and do not represent real numbers. For example, locations 1 through 7 represent different regions in the selected building stock, or the air conditioning (AC) class categories may be represented by 1 (no air conditioning), 2 (air-source heat pump), or 3 (window air conditioning). In the latter example, a building's AC would be represented by a value from the set $\{1,2,3\}$. The exact values of the categories depend on the chosen data set used for classification.

Response (r): The *response* data is the set of class values for one specific parameter, such as the building's surface area category or the location. Each building is represented by a single known value resulting in a vector of length b with values corresponding to the *class categories* for the parameter studied. Using the example from the *class categories* for air conditioning, the vector $[r_{AC}]$ of length b would contain air conditioning class values from the set $\{1,2,3\}$. These values are used to train the predictive model by establishing a link between the predictors (energy data) and responses (building parameters).

8 References

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.

Chalmers, Carl, William Hurst, Michael Mackay, and Paul Fergus. 2019. "Identifying Behavioural Changes for Health Monitoring Applications Using the Advanced Metering Infrastructure." *Behaviour & Information Technology* 38 (11). Taylor and Francis Ltd.: 1154–1166. doi:10.1080/0144929X.2019.1574900.

Djenouri, Djamel, Roufaida Laidi, Youcef Djenouri, and Ilangko Balasingham. 2019. "Machine Learning for Smart Building Applications." *ACM Computing Surveys (CSUR)* 52 (2). ACM PUB27 New York, NY, USA . doi:10.1145/3311950.

Esen, Hikmet, Mustafa Inalli, Abdulkadir Sengur, and Mehmet Esen. 2008. "Artificial Neural Networks and Adaptive Neuro-Fuzzy Assessments for Ground-Coupled Heat Pump System." *Energy and Buildings* 40 (6). Elsevier: 1074–1083.

doi:10.1016/J.ENBUILD.2007.10.002.

Fowler, K.M., A.H. Colotelo, J.L. Downs, K.D. Ham, J.W. Henderson, S.A. Montgmoery, S.A. Parker, and C.R. Vernon. 2015. *Simplified Processing Method for Meter Data Analysis*. Oak Ridge, TN.

Gianniou, Panagiota, Christoph Reinhart, David Hsu, Alfred Heller, and Carsten Rode. 2018. "Estimation of Temperature Setpoints and Heat Transfer Coefficients among Residential Buildings in Denmark Based on Smart Meter Data." *Building and Environment* 139 (July). Pergamon: 125–133. doi:10.1016/J.BUILDENV.2018.05.016.

Gładyszewska-Fiedoruk, Katarzyna, and Maria Jolanta Sulewska. 2020. "Thermal Comfort Evaluation Using Linear Discriminant Analysis (LDA) and Artificial Neural Networks (ANNs)." *Energies 2020, Vol. 13, Page 538* 13 (3). Multidisciplinary Digital Publishing Institute: 538. doi:10.3390/EN13030538.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 12th print. Springer-Verlag.

Himeur, Yassine, Khalida Ghanem, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. 2021. "Artificial Intelligence Based Anomaly Detection of Energy Consumption in Buildings: A Review, Current Trends and New Perspectives." *Applied Energy*. Elsevier Ltd. doi:10.1016/j.apenergy.2021.116601.

Hua, J., Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty. 2005. "Optimal Number of Features as a Function of Sample Size for Various Classification Rules." *Bioinformatics* 21 (8). Oxford Academic: 1509–1515. doi:10.1093/bioinformatics/bti171.

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

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning, with Applications in R*. Edited by Gareth M. James. 1st editio. Springer. doi:10.1007/978-1-4614-7138-7.

Klemenjak, Christoph. 2018. "On Performance Evaluation and Machine Learning Approaches in Non-Intrusive Load Monitoring." *Energy Informatics* 1 (S1). Springer Science and Business Media LLC: 36. doi:10.1186/s42162-018-0051-1.

Li, Dan, Guoqiang Hu, and Costas J. Spanos. 2016. "A Data-Driven Strategy for Detection and Diagnosis of Building Chiller Faults Using Linear Discriminant Analysis." *Energy and Buildings* 128 (September). Elsevier: 519–529. doi:10.1016/J.ENBUILD.2016.07.014.

Mathworks Inc. 2018. "Matlab Statistics and Machine Learning Toolbox R2018b." Natick, Massachusettes, United States.

Miller, Clayton, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W. Hobson, Zixiao Shi, and Forrest Meggers. 2020. "The Building Data Genome Project 2, Energy Meter Data from the ASHRAE Great Energy Predictor III Competition." *Scientific Data* 7 (1). Nature Research: 1–13. doi:10.1038/s41597-020-00712-x.

Mordor Intelligence. 2021. "Global Smart Meters Market | Growth, Trends, Forecasts (2020 - 2025)." https://www.mordorintelligence.com/industry-reports/global-smart-meters-market-industry.

Najafi, Behzad, Monica Depalo, Fabio Rinaldi, and Reza Arghandeh. 2021. "Building

Characterization through Smart Meter Data Analytics: Determination of the Most Influential Temporal and Importance-in-Prediction Based Features." *Energy and Buildings* 234 (March). Elsevier Ltd: 110671. doi:10.1016/j.enbuild.2020.110671.

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*, 3393–3400. Rome, Italy.

Neale, Adam, Michaël Kummert, and Michel Bernier. 2020a. "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. doi:10.1080/19401493.2020.1800096.

Neale, Adam, Michaël Kummert, and Michel Bernier. 2020b. "Virtual Smart Meter Data Set." https://vsmdata.meca.polymtl.ca/.

Oprea, Simona-Vasilica, Adela Bâra, Florina Camelia Puican, and Ioan Cosmin Radu. 2021. "Anomaly Detection with Machine Learning Algorithms and Big Data in Electricity Consumption." *Sustainability 2021, Vol. 13, Page 10963* 13 (19). Multidisciplinary Digital Publishing Institute: 10963. doi:10.3390/SU131910963.

Sarker, Iqbal H. 2021. "Machine Learning: Algorithms, Real-World Applications and Research Directions." *SN Computer Science* 2: 160. doi:10.1007/s42979-021-00592-x.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310. doi:10.1214/10-STS330.

Sokol, Julia, Carlos Cerezo Davila, Christoph F. Reinhart, C. Cerezo, and Christoph F.

Reinhart. 2016. “Validation of a Bayesian-Based Method for Defining Residential Archetypes in Urban Building Energy Models.” *Energy and Buildings* 134. Elsevier B.V.: 11–24.

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*. doi:10.1016/j.rser.2008.09.033.

Ullah, Amin, Kilichbek Haydarov, Ijaz Ul Haq, Khan Muhammad, Seungmin Rho, Miyoung Lee, and Sung Wook Baik. 2020. “Deep Learning Assisted Buildings Energy Consumption Profiling Using Smart Meter Data.” *Sensors 2020, Vol. 20, Page 873* 20 (3). Multidisciplinary Digital Publishing Institute: 873. doi:10.3390/S20030873.

Wang, Yi, Qixin Chen, Tao Hong, and Chongqing Kang. 2018. “Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges.” *IEEE Trans. Smart Grid*, 24. doi:10.1109/TSG.2018.2805.

Westermann, Paul, Chirag Deb, Arno Schlueter, and Ralph Evins. 2020. “Unsupervised Learning of Energy Signatures to Identify the Heating System and Building Type Using Smart Meter Data.” *Applied Energy* 264 (April). Elsevier: 14. doi:10.1016/J.APENERGY.2020.114715.

Zhang, Leping, Lu Wan, Yong Xiao, Shuangquan Li, and Chengpeng Zhu. 2019. “Anomaly Detection Method of Smart Meters Data Based on GMM-LDA Clustering Feature Learning and PSO Support Vector Machine.” In *ISPEC 2019 - 2019 IEEE Sustainable Power and Energy Conference: Grid Modernization for Energy Revolution, Proceedings*, 2407–2412. Institute of Electrical and Electronics

Engineers Inc. doi:10.1109/iSPEC48194.2019.8974989.

Zhang, Yang, Tao Huang, and Ettore Francesco Bompard. 2018. “Big Data Analytics in Smart Grids: A Review.” *Energy Informatics* 1 (1). Springer Science and Business Media LLC: 8. doi:10.1186/s42162-018-0007-5.

Appendix 1: Practical application of linear discriminant analysis

In order to illustrate a practical application of LDA, a data set of residential electricity consumption for the months of January and July for 1399 houses is presented in Figure A-1. In the data set there are 362 "small" houses (average area of 115 m^2) and 1037 "large" houses (average area of 250 m^2). A small set of data is used from the Virtual Smart Meter data set by Neale et al. (2020a) for the purpose of this appendix.

To summarize the example in the terms presented in the paper glossary:

- *Buildings*: 1399 single-family homes.
- *Features*: January and July electricity consumption. Each building is represented by two electricity consumption values $\{E_{jan}, E_{jul}\}$.
- *Class*: house size, represented by categories describing the size.
- *House size categories*: 'small' and 'large'.
- *Predictors*: the feature pairs $\{E_{jan}, E_{jul}\}$ for 1399 homes.
- *Response*: the size category labels for 1399 homes, either {'small'} or {'large'}.

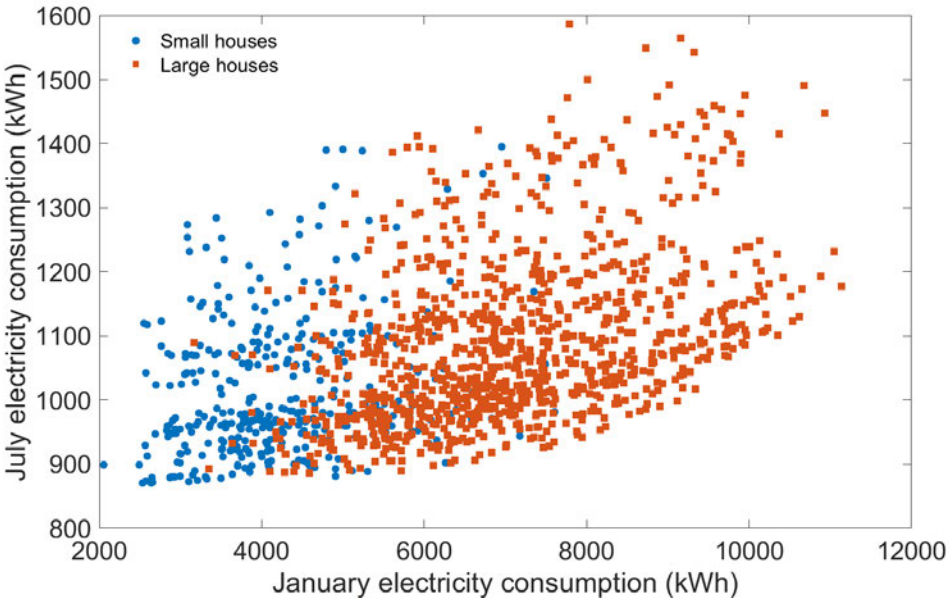

Figure A-1. January and July electricity consumption for small and large houses. Class: house size, categories (2): small and large, features (2): January and July electricity consumption.

The data in Figure A-1 illustrates that there is some degree of overlap between the small and large house data, which is due to the variety of building parameters used to model the homes. The goal of classification using LDA would be to establish a linear decision boundary that would best separate the two data sets such that new values of January and July electricity consumption will be classified as either "small" or "large". The probability p of a new data point belonging to one particular class category c can be expressed using Equation (A-1).

$$p_c = \frac{n_c}{n} \tag{A-1}$$

where n_c is number of samples in class category c, n is the total number of samples. For the example given, $p_{small} = 0.259$ and $p_{large} = 0.741$. Consider that the data can be divided into two subsets ($c = 2$), $\boldsymbol{Y}_{small}$ and $\boldsymbol{Y}_{large}$, which represent the electricity consumption data for the small houses and large houses, respectively, and can more generally be expressed as $\boldsymbol{Y}_c$. Each subset $\boldsymbol{Y}_c$ also has two features ($f = 2$) in this example, which can be expressed as the subsets $\boldsymbol{X}_{jan}$ and $\boldsymbol{X}_{jul}$, for January and July electricity consumption, respectively. Each subset can therefore be expressed as matrices $[\boldsymbol{X}_{small,jan}\ \boldsymbol{X}_{small,jul}]$ and $[\boldsymbol{X}_{large,jan}\ \boldsymbol{X}_{large,jul}]$ of size $[n_c \times f]$. Note that variables that are vectors or matrices are indicated with bold text.

The mean of each subset $\mathbf{X}$ can be calculated using Equation (A-2). Since there are a number of features per class category, the resulting mean values are stored in vector form of size $[1 \times f]$.

$$\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{i=1}^{f} \boldsymbol{X}_{c,i} \tag{A-2}$$

where $\boldsymbol{\mu}_c$ is a $[1 \times 2]$ vector containing the mean values of the January and July electricity consumption values for class category c, and i is the feature count. The mean

values can then be used to determine the within-class covariance, as expressed in Equation (A-3).

$$\boldsymbol{K}_c = \frac{1}{n_c - 1} \sum_{i=1}^{f} \left(\boldsymbol{X}_{c,i} - \boldsymbol{\mu}_c\right) \left(\boldsymbol{X}_{c,i} - \boldsymbol{\mu}_c\right)^T \tag{A-3}$$

where $\boldsymbol{K}_c$ is the covariance matrix for class category c of size $[f \times f]$, or $[2 \times 2]$ in this example. A common reduction technique used in LDA is to establish a pooled estimate of the covariance, combining the covariance matrices for the class categories. In the example given, for categories "small" and "large" houses, the pooled covariance could be expressed as in Equation (A-4).

$$\boldsymbol{K} = \frac{(n_{small} - 1)\boldsymbol{K}_{small} + (n_{small} - 1)\boldsymbol{K}_{large}}{n_{small} + n_{large} - 2} \tag{A-4}$$

where $\boldsymbol{K}$ is the pooled covariance matrix. By using the inverse of the covariance matrix, a discriminant function $\delta_c(x)$ can be determined, where the goal is to determine the maximum value of $\delta_c(x)$. For the purpose of brevity the derivation of Equation (A-5) is not presented here, but can be found in many reference texts related to machine learning techniques, such as in James et al. (2013).

$$\delta_c(x) = \boldsymbol{x}^T \boldsymbol{K}^{-1} \boldsymbol{\mu}_c - \frac{1}{2} \boldsymbol{\mu}_c^T \boldsymbol{K}^{-1} \boldsymbol{\mu}_c + log(p_c) \tag{A-5}$$

where $\boldsymbol{x}$ contains the feature variables, which in this case are the electricity consumption values in the months of January and July (E_{jan} and E_{jul}), and where the mean μ_c and the inverse matrix $\boldsymbol{K}^{-1}$ are determined using Equations (A-2) and (A-4), respectively. Equation (A-5) expresses the projection of the mean and covariance of the data sets on a projection axis and establishes a decision line by maximizing the term to the right of the equal sign. The process can be completed for each class category, as illustrated in Equations (A-6) and (A-7) for the *small house* and *large house* categories, respectively.

$$\delta_{small}(x) = [E_{jan} \quad E_{jul}]\boldsymbol{K}^{-1}\begin{bmatrix}\mu_{small,jan}\\ \mu_{small,jul}\end{bmatrix} - \frac{1}{2}[\mu_{small,jan} \quad \mu_{small,jul}]\boldsymbol{K}^{-1}\begin{bmatrix}\mu_{small,jan}\\ \mu_{small,jul}\end{bmatrix} + log(p_{small}) \tag{A-6}$$

$$\delta_{large}(x) = [E_{jan} \quad E_{jul}]\boldsymbol{K}^{-1}\begin{bmatrix}\mu_{large,jan}\\ \mu_{large,jul}\end{bmatrix} - \frac{1}{2}[\mu_{large,jan} \quad \mu_{large,jan}]\boldsymbol{K}^{-1}\begin{bmatrix}\mu_{large,jan}\\ \mu_{large,jul}\end{bmatrix} + log(p_{large}) \tag{A-7}$$

If the discriminant functions for each category are assumed equal ($\delta_{small}(x) = \delta_{large}(x)$), a linear decision boundary between those two categories can be established. By combining Equations (A-6) and (A-7) and simplifying, the resulting linear boundary separating the two class categories can be expressed as Equation (A-8).

$$h_{small:large} = \beta_{jan}E_{jan} + \beta_{jul}E_{jul} + C \tag{A-8}$$

where β_{jan}, β_{jul} and C are constants determined from the data and where $h_{small:large}$ is the classification rule result, which for a new coordinate of $\{E_{jan}, E_{jul}\}$ would determine whether the data point belongs to the first or second class category. If $h_{small:large} > 0$, the data belongs to the "small house" category. In this specific example there are only two categories, which indicates that if the data point is not a small house, it must be a large house, but in other cases there could be multiple other categories requiring additional classification boundaries to be verified. For the example given, the coefficients are described in Equation (A-9).

$$h_{small:large} = -0.00154E_{jan} + 0.00242E_{jul} + 5.6142 \tag{A-9}$$

The linear decision boundary in Equation (A-9) can be graphed by assuming $h_{small:large} = 0$, which is the line of transition between the two categories. By introducing the original data set values of $\{E_{jan}, E_{jul}\}$ into the classification boundary equation, the categories of the original data can be predicted and compared to the real

category values. For the studied example, the correct and incorrect predictions are illustrated in Figure A-2.

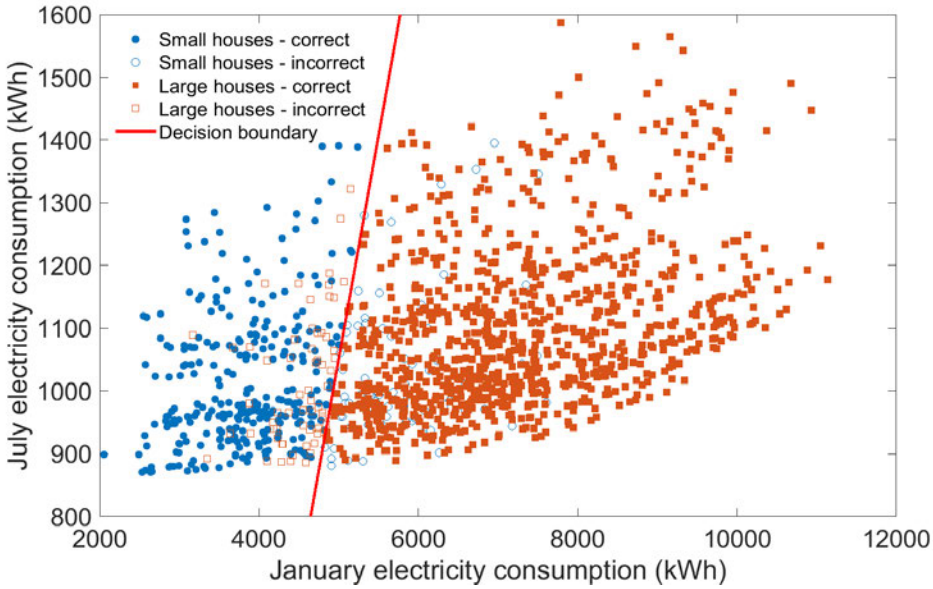

Figure A-2. Linear classification boundary with correct and incorrect predictions

For this case, the accuracy of the classification boundary can be expressed as the correct predictions divided by the number of data points, as described in Equation (A-10).

$$CA = \frac{CP}{TP} \tag{A-10}$$

where CA is the classification accuracy, CP is the number of correct predictions and TP is the total predictions for a validation set. For this study, k-fold cross-validation is used to determine CP for all predictive models, where $k = 5$. Readers unfamiliar with k-fold cross-validation may refer to the end of this appendix for a description of the method.

The example presented in Figure A-2 is simple in that it can be graphed in two dimensions, because there are only two features studied. With additional features the classification boundary becomes multidimensional, which can be expressed in a more general form as Equation (A-11).

$$h_{ci:cj} = C + \sum_{a=1}^{f} \beta_a E_a \tag{A-11}$$

where $h_{ci:cj}$ is the classification rule between class categories ci and cj, C is a constant and $\beta_a E_a$ are the corresponding coefficients and feature values. For smart meter data, the feature values are the electricity consumption values at various moments in time that depend on the desired time aggregation, i.e. hourly, daily, monthly, etc. Equation (A-11) requires the resolution of a matrix equation that scales exponentially with the number of features and class categories, which can rapidly become quite large considering a typical year of smart meter data recorded at 15-minute intervals has 35 040 data points.

Classification accuracy: *k*-fold cross-validation

The classification accuracy is determined by dividing the number of correct predictions (CP) by the number of total predictions (TP), as described in Equation (A-10). Establishing CP requires a data set to be divided into a test set and a validation set, commonly referred to as holdover validation. For holdover validation, the predictive model would be trained with some portion of the data set, such as 70%, and then the remaining data would be used to determine the classification accuracy. Since this prevents the use of the entire data set for training of the model, *k*-fold cross-validation is commonly used, which divides the data set into *k* segments and each segment is used to validate the predictive model developed with the remaining data, a process that is repeated *k* times. This allows the entire data set to be used for predictive model development and testing, but requires significantly longer computation time as the classification modeling is repeated multiple times. The classification accuracy for *k*-fold validation is expressed in Equation (A-12).

$$CA = \frac{\sum_{i=1}^{k_{folds}} CP_i}{TP} \tag{A-12}$$

where k_{folds} is the number of folds used for validation, which in the case of this study is 5-fold cross-validation, and CP_i is the number of correct predictions for fold i. There are 1260 correctly predicted data points out of the 1399 total data in the set. For the example illustrated in Figure A-2, CA is equal to 0.901, i.e. 90.1% of houses are correctly classified by the illustrated linear decision boundary. An illustration of 5-fold cross-validation is presented in Figure A-3.

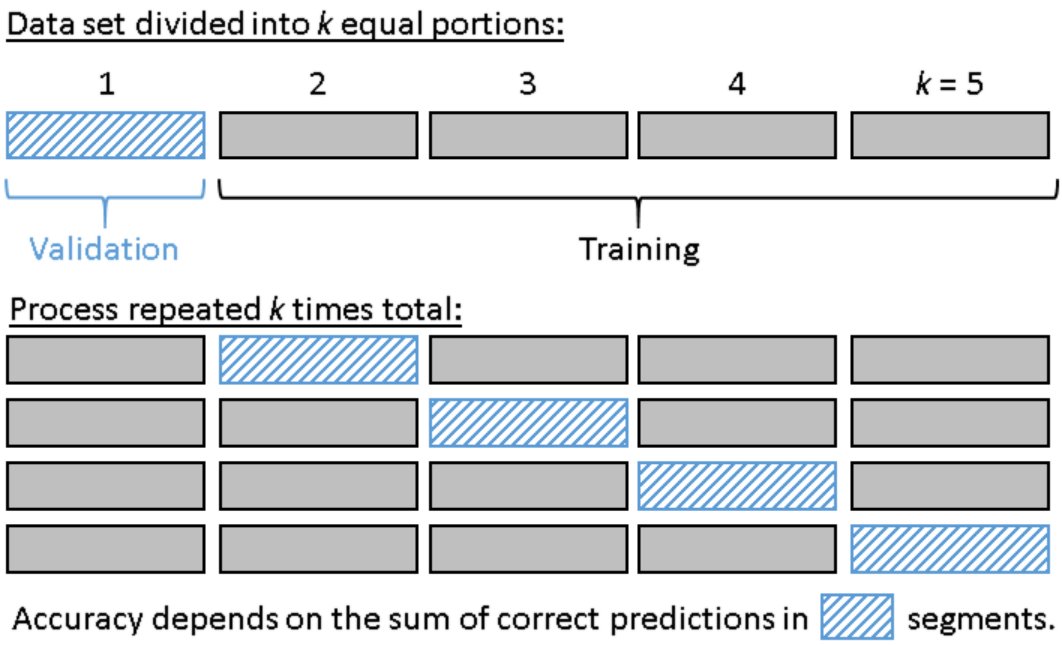

Figure A-3. Example of 5-fold cross-validation