

Residential Classification Using Limited Sample Size
Consumption Data with Associated Metadata

by

Chunjin Liu

Bachelor of Electrical Engineering, Beijing Jiaotong University, 2007
Master of Power Electronics and Power Drive, Beijing Jiaotong University, 2009

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master of Science in Engineering

in the Graduate Academic Unit of Electrical and Computer Engineering

Supervisor: Julian Meng, PhD, Electrical and Computer Engineering

Examining Board: Saleh A. Saleh, PhD, Electrical and Computer Engineering, Chair
Eugene F. Hill, PhD, Electrical and Computer Engineering
Eric D. Hildebrand, PhD, Civil Engineering

This thesis is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

February 2023

©Chunjin Liu 2023

ABSTRACT

The ongoing deployment of residential smart meters in numerous jurisdictions will inevitably lead to a significant amount of electricity consumption data becoming available. This information presents a valuable opportunity to utilities for mining and analysis of load patterns. Household load shapes can reveal significant differences among large groups of households in the magnitude and timing of their electricity consumption. However, limitations in dataset sample size and the associated descriptive data (i.e. metadata) makes it difficult for residential classification problems. It is often desirable to classify residential customers based on their energy consumption profiles and give meaningful insight into the amount of energy these customers use over a specified period. The research work presented in this thesis focusses on specific classification requirements of two limited sample size datasets which have differing metadata associated with each. For Dataset 1, clustering techniques, statistical analysis and machine learning (ML) algorithms were used to analyze the difference between residential consumption profiles based only on lot size and home price. Strategic groupings such as more expensive homes on large lots versus lower cost homes on small lots have shown to have some separability in terms of load profiles. Dataset 2 has only heating source type associated with it and the main focus was to differentiate homes that utilize one of three heating types: electric baseboard (BB), air-source heat pumps (HP) or mini-split heat pump (MS). This will help provide information on expected consumption patterns for the presence of a given heating type. Statistical and ML techniques were again applied to this problem and performance is assessed using numerous load profile and weather features.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Julian Meng for his guidance and knowledge throughout my research work. Also, I would like to thank Dr. Julian Cardenas and Dr. Eugene Hill for their advice and suggestions for this research work. Thank you to NB Power and Berend van Middeldorp for providing the data for this thesis, funding of this research work, and the opportunity to explore such a unique topic.

Finally, I would like to thank my family for their support.

Table of Contents

ABSTRACT	ii
ACKNOWLEDGEMENTS	iii
Table of Contents	iv
List of Tables	vii
List of Figures	viii
List of Acronyms	x
1. Introduction	1
2. Statistical techniques and ML algorithms	12
2.1 Background	12
2.2 Peak-to-mean ratio (PMR).....	12
2.3 K-means clustering for load profile grouping [3] [39]-[42]	13
2.3.1 General.....	13
2.3.2 Representative daily load shapes	14
2.4 Student's T-test [3] [43]-[47].....	15
2.4.1 General.....	15
2.4.2 T-test mathematical formulation.....	16
2.5 Machine learning based classification algorithms	18
2.5.1 Random Forest algorithm [48] [49].....	18
2.5.2 Support vector machine algorithm [50]-[54]	20
2.5.3 k-nearest neighbors algorithm [55] [56]	22

2.5.4	Predictor/Feature importance [57]	23
2.5.5	Evaluation Metrics [37][58][59]	24
3.	Dataset 1: Residential classification by value, lot size and type	25
3.1	General.....	25
3.2	Data analysis based on peak-to-mean ratio.....	26
3.3	Grouping data based on combined metadata features	28
3.3.1	The assessment of combining metadata features	28
3.3.2	Representative daily load shapes based on K-means clustering	30
3.3.3	Student's T-test	32
3.3.4	Monthly T-test analysis	33
3.4	Classification based on ML algorithms	35
3.5	Analysis and summary.....	39
4	Dataset 2: Residential heat source classification.....	41
4.1	General.....	41
4.2	Grouping data by heating source.....	42
4.2.1	Heating source analysis	42
4.2.2	Load shape analysis: T-test results.....	45
4.3	Load profile variations of electric space heating types	47
4.3.1	HP vs BB	48
4.3.2	MS vs BB.....	50
4.3.3	MS vs HP	52

4.4	Electric heating source classification based on ML algorithms.....	54
4.5	Analysis and summary.....	59
5	Conclusion.....	61
5.1	Further comments of Dataset 1	62
5.2	Further comments of Dataset 2	63
5.3	Future work.....	63
	Bibliography	65
	Appendix A	70
	Appendix B	71
	Appendix C	73
	Curriculum Vitae	

List of Tables

Table 3-1 Dataset 1 of NB Power residences based on available metadata	25
Table 3-2 T-test results 2018	33
Table 3-3 T-test results in 2018.....	34
Table 3-4 Consumption data classification results.....	37
Table 4-1 Dataset 2 of NB Power residences based on heating source.....	42
Table 4-2 Dataset 2: Average energy consumption in each month.....	44
Table 4-3 T-test results in 2020.....	45
Table 4-4 T-test result (HP vs NE).....	46
Table 4-5 T-test result (MS vs NE heating).....	46
Table 4-6 T-test result (HP vs BB).....	50
Table 4-7 T-test result (MS vs BB)	52
Table 4-8 T-test results (MS vs HP).....	54
Table 4-9 Consumption data classification results.....	56

List of Figures

Fig. 1-1 2018/19 New Brunswick Energy sales by customer classification.....	1
Fig. 1-2 Approximate average end use breakdown of residential electricity consumption in 2018/19	2
Fig. 2-1 Demonstration of K-means clustering algorithm [41].....	13
Fig. 2-2 Two-tailed test (assuming 5% significance level, split 2.5% each on either side)	16
Fig. 2-3 The schematic diagram of random forest [49]	19
Fig. 2-4 The demonstration of SVM in 2-D space [51].....	21
Fig. 2-5 The demonstration of KNN	22
Fig. 3-1 Scatter plots of grouping results based on PMR metric	27
Fig. 3-2 Peak power and mean power distribution for combining features.....	29
Fig. 3-3 Two feature metadata grouping results based on K-means clustering	31
Fig. 3-4 Two feature metadata grouping T-test results	33
Fig. 3-5 Supervised learning load classification results	36
Fig. 3-6 True positive rate and false negative rate of different subset	37
Fig. 3-7 Positive predictive value and false discovery rate of different subset	38
Fig. 4-1 Average daily load shapes for different heating sources	43
Fig. 4-2 Representative daily load shapes	47
Fig. 4-3 Peak power and mean power (HP vs BB)	48
Fig. 4-4 # of load shapes in each cluster (HP vs BB).....	49
Fig. 4-5 T-test results (HP vs BB).....	49
Fig. 4-6 Peak power and mean power (MS vs BB).....	50

Fig. 4-7 # of load shapes in each cluster.....	51
Fig. 4-8 T-test results (MS vs BB)	51
Fig. 4-9 Peak power and mean power (MS vs HP).....	52
Fig. 4-10 # of load shapes in each cluster.....	53
Fig. 4-11 T-test results (MS vs HP).....	53
Fig. 4-12 Supervised learning load classification results	56
Fig. 4-13 True positive rate and false negative rate of different groups	57
Fig. 4-14 Positive predictive value and false discovery rate of different subset.....	58

List of Acronyms

- AMI** Advanced metering infrastructure
- ANN** Artificial neural network
- AUC** Area under the ROC curve
- BB** Baseboard
- BSVM** Biased support vector machine
- BTU** British thermal unit
- CAFD** Characteristic attributes in the frequency domain
- CART** Classification and regression tree
- CCA** Curvilinear component analysis
- CCF** Cluster classify forecast
- CVMM** C-vine copulas mixture model
- DNN** Deep neural network
- DR** Demand response
- DT** Decision tree
- EE** Energy efficiency
- ELR** Enhanced logistic regression
- FDR** False discovery rate
- FNR** False negative rate
- FWK** Functional wavelet-kernel approach
- HMM** Hidden Markov model
- HP** Air-source heat pumps
- IBDR** Incentive-based demand response

KNN k-nearest neighbor

MI Mutual information

ML Machine learning

MS Mini-split heat pump

NE Non-electric

PCA Principal component analysis

PMR Peak-to-mean ratio

PPV Positive predictive value

RBF Radial basis function

RFE Recursive feature elimination

RF Random forest

ROC Receiver operating characteristic

RVMM R-vine copula mixture model

SMEs Small and medium businesses

SVM Support vector machine

TPR True positive rate

UMAP Uniform manifold approximation and projection

CHAPTER 1

1. INTRODUCTION

1.1 General

NB Power's in-province electrical loads are divided into three main groups: residential, general service, and industrial. In the fiscal year 2018/19, residential customers accounted for 44.5 percent of the total in-province electrical energy sales (40 per cent directly by NB Power and 4.5 percent by Wholesale utilities). The residential classification is made up mostly of year-round domestic (household) customers. Growth in the residential sector is mainly due to population and customer growth as well as increasing penetration of electric heat [1].

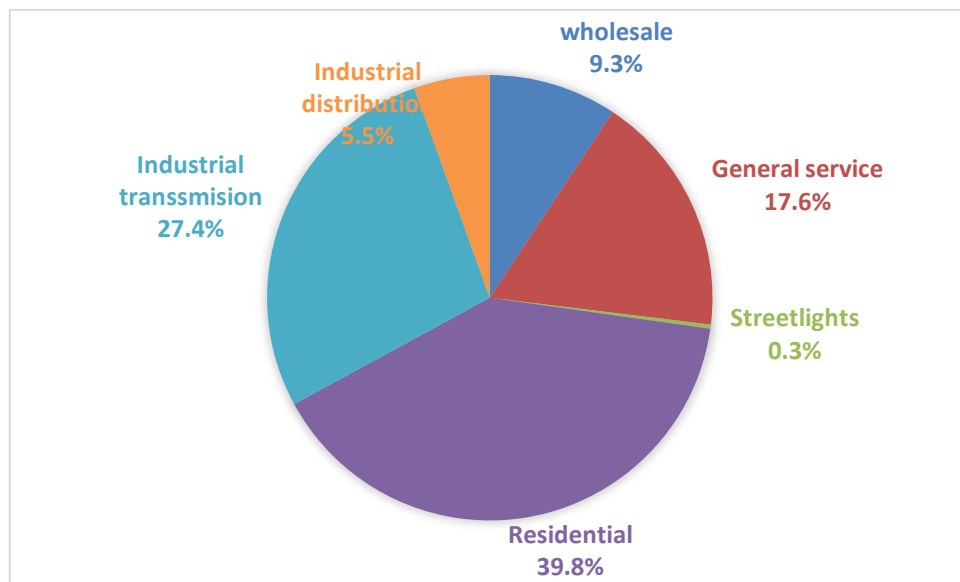


Fig. 1-1 2018/19 New Brunswick Energy sales by customer classification

The NB average household energy is generally comprised of electric space heating, domestic hot water heating and other uses. In 2018/19 the average household energy was comprised of approximately 48 percent electric space heating, 17 percent domestic hot

water heating and 36 percent other uses. The breakdown of energy consumption is shown in Figure 1-2.

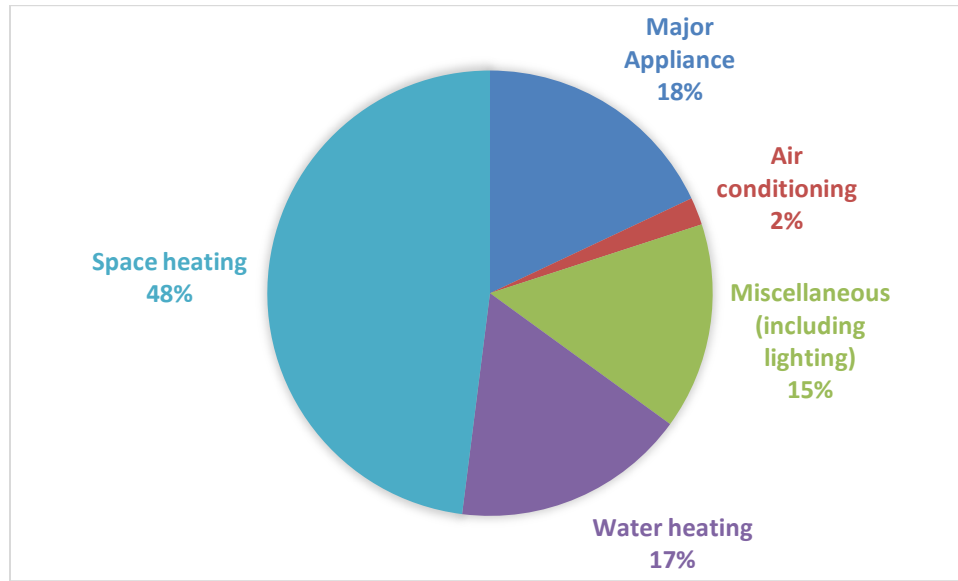


Fig. 1-2 Approximate average end use breakdown of residential electricity consumption in 2018/19

The mixture of household energy can change over time as a result of a number of factors including changes in electric space and water heating technologies, as well as appliance trends and efficiency standards. To account for such trends, the energy forecast for the residential class is based on an end use model that requires identification of the various applications of electricity. The penetration level and use for each type of household application can provide insight for the average use per customer and their given load profile. This combined with the number of customers leads to an estimate of the aggregate load from residential customers. In 2018/19 there were 349,000 year-round residential customers in New Brunswick. Of those, 309,000 were served directly by NB Power and 40,000 were served by the municipal utilities in the cities of Saint John and Edmundston and the town of Perth-Andover [1].

The deployment of a residential advanced metering infrastructure (AMI) using smart meters has made detailed energy consumption data available for domestic (household) customers [1]. The wide-spread dissemination of electricity smart meters offers the opportunity to create load profile assessment strategies based on 15 min, 30 min, or hourly household energy usage. Understanding a household's time of day energy consumption, daily usage pattern stability over time, as well as actual volume of energy usage, offers insights into household use of energy. Household load shapes can reveal significant differences among large groups of households in the magnitude and timing of their electricity consumption [2]. Further, this consumption lifestyle information has the potential to enhance targeting and tailoring of demand response (DR) and energy efficiency (EE) programs as well as promoting energy reduction recommendations [3]. The number of research studies in energy consumption, including annual consumption for various uses and the characteristics impacting energy usage and consumption prediction, is dramatically increasing with the deployment of smart meters and other data collection methods. Therefore, determining residential characteristics based on usage patterns is important to utilities for preliminary studies in assessing future energy needs and for grid planning.

1.2 Problem statement

Although AMI deployment has clear advantages, some jurisdictions are slower to adapt newer technologies thereby limiting the availability of residential energy consumption data and any associated metadata. Here, the term metadata [4] is introduced to describe "data about data" and is defined as the data providing information about one or more aspects of the measured data. It is used to summarize basic information about data which can make tracking and working with specific data easier. In this regard, data quality

is dependent on metadata such as home size, price, heating type, location, occupancy etc. and is important in the prediction of residential characteristics from a daily usage pattern. Techniques based on clustering and machine learning (ML) require key metadata features to enable residential groupings to be clearly delineated based on energy consumption.

1.3 Thesis objectives

The main research objective of this research is to group residential customers based on residential consumption data with limited sample size consumption data and associated metadata, and to provide meaningful information into the amount of energy these customers will use over a specified period of time.

Research objectives for this thesis are separated based on the analysis of two different residential energy consumption datasets depending on the metadata associated with it:

- **Dataset 1:** 15-minute interval residential consumption data for 103 homes (provided by NB Power) for the year 2018 which has limited metadata of home price, lot size, and type. It does not contain any information on four key features such as home size, heating type, occupancy and location. For this dataset, various methods were used to determine statistical differences in residential load profiles based on the available metadata. Basic consumption metrics, unsupervised K-means clustering and supervised ML techniques were evaluated. For the latter, three ML algorithms were assessed: random forest (RF), support vector machine (SVM) and k-nearest neighbor (KNN).
- **Dataset 2:** Another 15-minute interval residential dataset approximately 500 homes (provided by NB Power from 2020) has only the metadata of heating type(s) within the home and the home's general location. Along with some statistical

analysis, various ML methods were implemented to determine whether a home has one of three heating types: electric baseboard (BB), air-source heat pumps (HP) or mini-split heat pump (MS). Important to this work was the utilization of a large number of consumption and weather features during the ML training stages and determining performance impacts with reducing this number.

1.4 Literature review

Energy consumption has historically been of significant interest for economics and engineering analysis. In the smart grid era, smart meter data is playing a vital role. With the recent large-scale introduction of smart meters, many studies have been published that address high resolution time series modeling or customer clustering. The key challenge for household load forecasting lies in the high volatility and uncertainty of load profiles [5]. The load data is highly volatile due to its variability in load pattern. The load pattern of every customer varies with respect to certain factors such as demographic information, environmental factors, time stamp, etc. [6]

A load survey study was performed to find the typical load patterns of different customer classes [5]. The power consumption of customers was collected, and statistical methods were applied to derive the typical load patterns of each customer class. The study in [6] proposes a novel method, a dynamic synthesis load modeling approach, which is based on load survey and curve analysis techniques. Utilizing various kinds of field data, this approach can create a load model reflecting its time-varying property and also provides a new aspect of load modeling in the modern power grid.

Article [7] analyzes household electricity consumption in Estonia. Energy consumption of workday and holiday by load type is discussed. In [8] an approach allowing

incomplete load profiles to be classified while maintaining a less than 10% classification error with up to 20% of the data missing was presented. The study in [9] analyzes the energy consumption at a single household level using smart meter data to improve residential energy services and gain insights into planning demand response programs. The analysis demonstrates that forecasting residential energy consumption for individual households is feasible, but the accuracy is highly dependent on household behavior variability.

Classification of loads is their division into sub-groups based on specific criteria. The research in [3] investigates a household electricity classification methodology that uses an encoding system with a pre-processed load shape dictionary. K-means clustering, adaptive K-means clustering, and hierarchical clustering were used for the classification. An approach for clustering residential end-users based on their load characteristics is presented in [10]. The methodology can reduce hourly load data with a stratification approach, together with K-means which clusters the customers based on their relative load levels. The work in [11] investigates a new load profile classification approach based on adaptive K-means clustering over an energy consumption dataset. The proposed method identifies dominant patterns for each house and also examines the seasonal impact. In [12], load profile data is analyzed for getting demand patterns of customers, and an algorithm for clustering similar patterns is developed to generate representative curves. However, most of the data are for commercial and industrial loads. The classified demand patterns are not used for analyzing customer information, actual load behaviors, tariff design, load forecasting, and so on. Self-organizing maps (SOM) and K-means are used to find load patterns in [13]. The study in [14] illustrates and compares the results obtained by using

various unsupervised clustering algorithms (modified follow-the-leader, hierarchical clustering, K-means, fuzzy K-means) and self-organizing maps to group together customers with similar electrical behavior. Furthermore, it discusses and compares various techniques (Sammon map, principal component analysis (PCA), and curvilinear component analysis (CCA)) to allow for storing a relatively small amount of data in the database of the distribution service provider for customer classification purposes. The analysis in [15] uses adaptive K-means clustering for load profile classification of residential energy consumption.

In [2], it is proposed to infer occupancy states from consumption time series data using a Hidden Markov Model (HMM) framework. It shows that users may be classified into groups according to their consumption patterns which exhibit qualitatively different dynamics. It investigates empirically the information that residential energy consumers' temporal energy demand patterns may convey about their demographic, household, and appliance characteristics. In [16], different power consumption scenarios using the HMM method of pattern recognition are assessed for a variety of user-appliance combinations: single-user, single-appliance; single-user, multi-appliances; multi-users, multi-appliances usage. It also provides an in-depth analysis of the variability of such use cases.

An adaptive fuzzy C-means algorithm is used to cluster the residents using workday load data and holiday load data in [17]. A load characteristics analysis is carried out, and a new user classification method is proposed to realize the differences among the residential electricity behavior characteristics. Article [18] presents a self-organization based integrated model for customer classification and load profiling in distribution systems. The results demonstrate that the methodology is able to be efficiently used in distribution

systems where the supplied customer information is very poor and is based only on the data provided by classic meters.

A novel methodology, functional wavelet-kernel approach (FWK), able to improve short term functional time series forecasts of household-level electricity demand has been introduced in [19]. An unsupervised classification method was adopted to find similar segments.

Copulas are used to describe, or model, the dependence among random variables. Paper [20] presents a novel finite mixture modeling framework based on C-vine copulas (CVMM) for carrying out consumer categorization. The superiority of the proposed framework lies in the great flexibility of pair copulas in being able to identify multidimensional dependency structures present in load profiling data.

A study in [21] presents a complete framework including a clustering module and a classification module for load pattern identification. Firstly, an innovative mixture model based on R-vine copula (RVMM) is proposed for clustering load profiling data and obtaining typical load patterns. Then, a RF model is constructed with certain load characteristic indexes, and a RVMM clustering result is employed as a supervised classification model to predict the category of new customers.

In [22] a novel electricity load forecasting architecture is developed that integrates three modules into a single model: RF and recursive feature elimination methods for data selection, T-stochastic neighborhood embedding algorithm for extraction, and deep neural network (DNN) for classification. The research in [23] proposes a deep learning-based load forecasting method that uses aggregated smart meter data along with the demographic

information. The performance of DNN is compared with a shallow network and was found to outperform it in terms of error calculations.

To directly learn the uncertainty of load profiles, a novel pooling-based deep recurrent neural network is proposed in [24], which batches a group of customers' load profiles into a pool of inputs.

In [25], an accurate electricity load and price forecasting model has been proposed, which consists of feature engineering and classification. To remove irrelevant features, decision tree (DT) and recursive feature elimination (RFE) methods are used. Features are extracted through mutual information (MI) after removing uncertainty. To attain accurate electricity load and price forecasting, an enhanced logistic regression (ELR) classifier is proposed.

The work presented in [26] attempts to formulate the theoretical framework for customer classification using the annual load profiles. It demonstrates how to extract characteristic attributes in the frequency domain (CAFD) and to use these CAFDs to formulate a hierarchy of load profiles that can be used as the systematic framework for customer load classification. As signatures for customer classes and subclasses, the CAFDs are obtained by using a data mining method called CART (classification and regression tree).

The study in [27] proposes a novel method of cluster-classify-forecast (CCF) for individual household electricity load forecasts. It integrates the information contained in typical daily consumption profiles extracted by clustering and classification methods. In addition to the improvements in forecast performance, the method reveals key information

about a household's habitual load profiles and other important variables which impact household consumption.

Article [28] shows how temporal resolution of power demand profiles affects the quality of the clustering process, the consistency of cluster membership (profiles exhibiting similar behavior), and the efficiency of the clustering process.

In [29], a residential load classification algorithm based on multi-model parallel integration (MMPI) is proposed. The classification results show that it has higher classification accuracy and efficiency, compared with the traditional single model load classification.

In a study in [30], the uniform manifold approximation and projection (UMAP) is used to compress electricity consumption data. The RF classification algorithm is then used on the compressed data to learn the consumption patterns of two distinctive user bases - household consumers and SMEs (small and medium businesses).

A novel customer baseline load (CBL) method for residential customers that uses the data reconstruction capability of a stacked autoencoder (SAE) is described in [31]. In the model, two SAEs are synchronously trained—one SAE generates a pseudo-load pool and the second one is used to select a pseudo-load to reconstruct a residential CBL. A SVM classifier is self-trained to conduct the pseudo-load selection.

To classify energy load curves according to their similarity with other households, the entropy as a quantitative metric for typical load curve classification and clustering is presented in [32]. It also introduces the likelihood of time periods to uniquely distinguish load curves of residential households and approximate the minimal required time resolution for classification tasks.

The research in [33], an incentive-based demand response (IBDR) is proposed to segregate the elasticity of household appliances. The proposed elasticity approach provides a more accurate model of both the load shift and load reduction potential in the residential sector. Customer classification is explored to understand the diversity of customer behavior. In [34], a deep dictionary learning and deep transform learning based multi-label classification technique is introduced to identify the classes given the aggregate smart-meter reading in order to accomplish non-intrusive appliance load monitoring.

In [35], it shows that utility companies can extract household-specific information from smart meter data using machine learning. Features from smart meter and weather data are derived by [36] and the RF classifier is used to recognize household classes. The results indicate that even datasets with an hourly or daily resolution are sufficient to impute key household characteristics with decent accuracy. The study in [37] investigates the prediction of heat pump installations, their thermal reservoir and age. A dataset with 397 smart meter households in Switzerland is obtained to collected ground truth data on installed heat pumps, weather data and geographical information., ML algorithms such as RF, SVM, KNN and artificial neural network (ANN) are used and compared to detect the existence of heat pumps, the heat reservoir, and age with a high area under the receiver operating characteristic (ROC) curve (AUC) performance metric. The paper [38] aims to detect electric heat pumps from coarse grained smart meter data for a heat pump marketing campaign. Features highly relevant to heat pump usage from smart meter data and weather data are extracted, and biased support vector machine (BSVM) is applied to the implement the load classification.

CHAPTER 2

2. STATISTICAL TECHNIQUES AND ML ALGORITHMS

2.1 Background

This chapter describes some foundational aspects of assessing the residential consumption data based on basic metrics and more complex data analysis techniques. Analysis is based on simple metrics based on daily consumption patterns and then becoming more complex to assess load profile groupings based on K-means clustering with statistical analyses to test the null hypothesis on group differences. This is followed by a discussion of ML techniques used different residential group classifications and the evaluation criteria used for the purposes of assessing the testing accuracy. In this project, MATLAB and RStudio are used to implement all metric calculations, feature extraction, classification, and analytical methods.

2.2 Peak-to-mean ratio (PMR)

This metric investigates whether basic load demand metrics such as peak load and average daily consumption are able to distinguish or group residential users based on available metadata features. The process is as follows:

- (1) For each household, identify the mean daily consumption of a month.
- (2) Identify the peak consumption and calculate the ratio of peak-power to mean-power (i.e PMR).
- (3) Identify the time of day at which that peak consumption occurs.

As an example, the metrics above are assessed to see if they can separate home consumption profiles based on a single metadata feature such as lot size or home price. This assessment is especially important in the colder and warmer months of the year.

2.3 K-means clustering for load profile grouping [3] [39]-[42]

This method is used to determine statistically significant load profiles based on selected metadata features and is based on a variation of the well-known unsupervised K-means algorithm. This method utilizes clustering methodology to implement shape classification, followed by a statistical analysis to determine load profiles are unique to specific residential groups. The Student's T-test is used to assess the statistical significance of load profiles differing in shape for different clusters. Other clustering algorithms can also be used to segment customers with similar consumption behavior, such as: K-means, adaptive K-means, hierarchical clustering, fuzzy K-means, SOM, follow-the-leader, HMM, fuzzy C-means etc.

2.3.1 General

K-means clustering is one of the most popular unsupervised machine learning algorithms. This algorithm aims to partition N observations into K ($K \leq N$) clusters. Each observation belongs to the cluster with the nearest mean (cluster center or cluster centroid) serving as a prototype of the cluster.

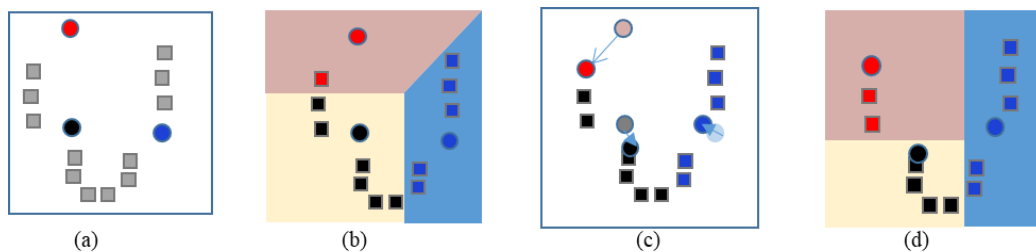


Fig. 2-1 Demonstration of K-means clustering algorithm [41]

The demonstration of the K-means clustering algorithm is shown in Fig. 2-1 for a 2-D spatial clustering. To process the learning data, the number of clusters K needs to have a predetermined value before using the algorithm. The K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning

points for every cluster. It then performs iterative (repetitive) calculations to optimize the positions of the centroids. The algorithm steps are shown below for a case where $K=3$:

- (1) Fig.2-1(a): The K "initial means" which are represented by dots of different colors, are randomly generated within the data domain, and the gray squares represent N observations.
- (2) Fig.2-1(b): K clusters are created by associating every observation with the nearest mean.
- (3) Fig.2-1(c): The centroid of each of the K clusters becomes the new mean.
- (4) Fig.2-1(d): Steps 2 and 3 are repeated until convergence has been reached.

In our research, the clustering will be performed in a 96-dimensional space given daily 15-minute interval consumption data.

2.3.2 Representative daily load shapes

This algorithm also applies to multi-dimensional vectors. Given a set of observations $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$, where each observation is a D -dimensional real vector ($D = 96$), K -means clustering aims to partition the N observations (or number of residential load profiles) into K sets $\mathbf{S} = \{S_1, S_2, \dots, S_K\}$ so as to minimize the within-cluster variances based on some metric such as squared Euclidean distances,

$$\arg \min_{\mathbf{S}} \sum_{i=1}^K \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mathbf{C}_i\|^2 \quad (2-1)$$

where \mathbf{C}_i is the cluster centroid in S_i .

For daily energy consumption, the clustering data contains K representative load shapes (cluster centroids) $\mathbf{C}_i(t)$, and every load shape in the data is mapped to the closest

cluster centroid. Load shape $x(t)$ is assigned to center $i^*(x)$ that minimizes the squared error:

$$E(x, i) = \sum_{t=1}^D (C_i(t) - x(t))^2 \quad (2-2)$$

$$i^*(x) = \arg \min_i E(x, i) \quad (2-3)$$

Cluster centroids can generally be determined in a supervised or unsupervised manner. In this research, the latter is used initially to determine the separability of residential energy consumption based on residential load patterns. Euclidean distance is then used to determine load profile groupings.

2.4 Student's T-test [3] [43]-[47]

Once K-means clustering is used to create the representative daily load shapes library, to compare the distribution of load shapes in different clusters, a two-tailed Student's T-test is introduced to test whether groups based on a particular metadata feature or groupings of metafeatures have statistically different load profiles.

2.4.1 General

The T-test is a statistical hypothesis test in which the test statistic follows a Student's T-distribution under the null hypothesis. The T-test tells how significant the differences between group means are. The T-test is usually used when the sample data follows a normal distribution and the population variance is unknown.

In this research, a two-sided (also termed two-tailed) T-test is introduced to specify whether the observed sample is larger or smaller than the hypothesized certain range of values. It is used in null-hypothesis testing and testing for statistical differences. For the Student T-test, the test statistic T is calculated and compared to the critical value to accept

or reject the null hypothesis (i.e there is no statistical difference). The appropriate reference distribution for the T-statistic is the T-distribution and the critical value depends on the significance level of the test (the probability of erroneously rejecting the null hypothesis). A two tailed test of 5% significance level is shown in Fig. 2-2.

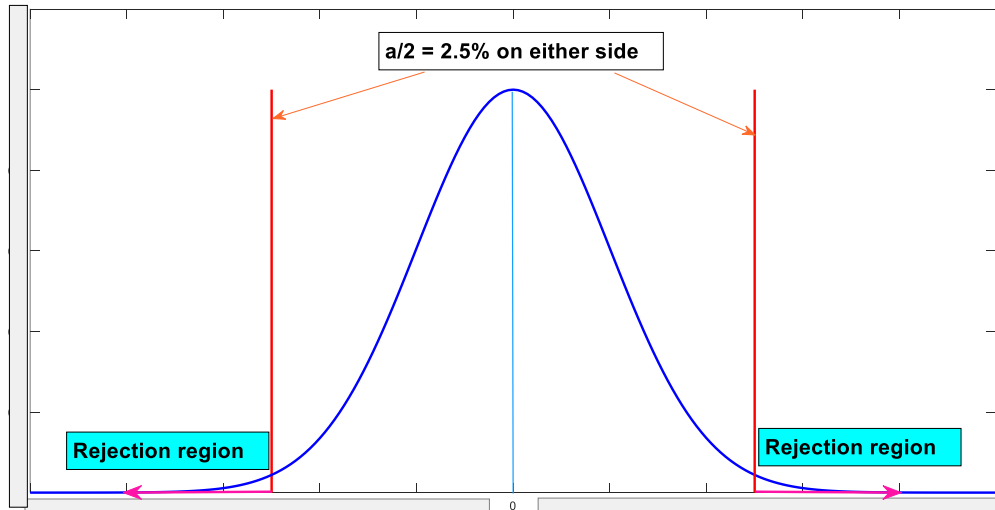


Fig. 2-2 Two-tailed test (assuming 5% significance level, split 2.5% each on either side)

2.4.2 T-test mathematical formulation

The T-test can be used to determine if the means of two sets of data are significantly different from each other on a probabilistic basis. In this research, the general steps of the T-test are used check $P(C_i | \text{condition A}) = P(C_i | \text{condition B})$ are shown below:

N_1 : sample size satisfying condition A

N_2 : sample size satisfying condition B

$$\bar{X}_1: \frac{\# \text{ of } C_i \text{ among } N_1}{N_1}, \quad \bar{X}_2: \frac{\# \text{ of } C_i \text{ among } N_2}{N_2}$$

$$S_1^2 = \bar{X}_1(1 - \bar{X}_1) \quad S_2^2 = \bar{X}_2(1 - \bar{X}_2) \quad (2-4)$$

The T-statistic to test whether the means are different can be calculated as follows:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{N_1} + \frac{1}{N_2}} \sqrt{\frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}}}, \text{ d.f.} = N_1 + N_2 - 2 \quad (2-5)$$

where the following three options are considered:

- 1) $T < t_{0.025} : P(C_i | \text{condition A}) < P(C_i | \text{condition B})$,
- 2) $T > t_{0.975} : P(C_i | \text{condition A}) > P(C_i | \text{condition B})$,
- 3) Otherwise: $P(C_i | \text{condition A}) = P(C_i | \text{condition B})$.

and

$$s_p = \sqrt{\frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}} \quad (2-6)$$

is the pooled standard deviation of the two samples. It is defined in this way so that its square is an unbiased estimator of the common variance whether or not the population means are the same. In these formulas, $N_i - 1$ is the number of degrees of freedom for each group, and the total sample size minus two (that is $N_1 + N_2 - 2$) is the total number of degrees of freedom (d.f.), which is used in significance testing. For our research, C_i are 96-dimensional cluster centroids generate from daily consumption data.

In addition, a T-test uses a T-statistic and compares this to T-distribution values to determine if the results are statistically significant. T-distribution values are shown in the Appendix A [47]. In this research, T- distribution values of two-tailed 5% significance level are chosen, splitting 2.5% each on either side. According to the d.f., the T-distribution critical values, $t_{0.025}$ and $t_{0.975}$ can be acquired using table look-up.

2.5 Machine learning based classification algorithms

Since heating type is very important for load profiles, Dataset 2 was analyzed to determine if heating type alone is useful in grouping residential customers. It is argued that this is necessary information as the load shapes are highly dependent on the nature of how the home is heated or cooled. By means of supervised ML, it is intended to discover individual characteristics of the dominant electric heating types used in NB. An important part of this study is to evaluate predictor or feature importance for the ML training process. These features are based on daily consumption and weather patterns.

Feature vectors were derived from consumption data collected at 15-min granularity over one week (7 days) and from weather data at 30-min intervals over one week. The package of “SmartMeterAnalytics” in RStudio is used to extract the features of smart meter data and weather data [35] [36]. Supervised ML algorithms such as RF, SVM, and KNN are used to classify the load profiles of different electrical heating types [37] [38]. The performance of these algorithms is evaluated and analyzed.

2.5.1 Random Forest algorithm [48] [49]

Existing literature on analysis of smart meter data focuses mainly on forecasting and load classification. Many classification algorithms are used to categorize the customer consumption data, such as DNN, DT and RFE, RF etc. There are many mixtures or improved algorithms introduced to deal with the consumption data load profiles.

RF, as its name implies, consists of a large number of individual decision trees that operate as an ensemble [48]. The fundamental concept behind RF is a simple but powerful one -- the wisdom of crowds. A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. Each

individual tree in the RF outputs a class prediction and the class with the most votes becomes the model's prediction. RF allows each individual tree to randomly sample from the dataset with replacement, resulting in different trees. This process is known as bagging (Bootstrap Aggregation), which ensure that the behavior of each individual tree is not too correlated with the behavior of any of the other trees in the model.

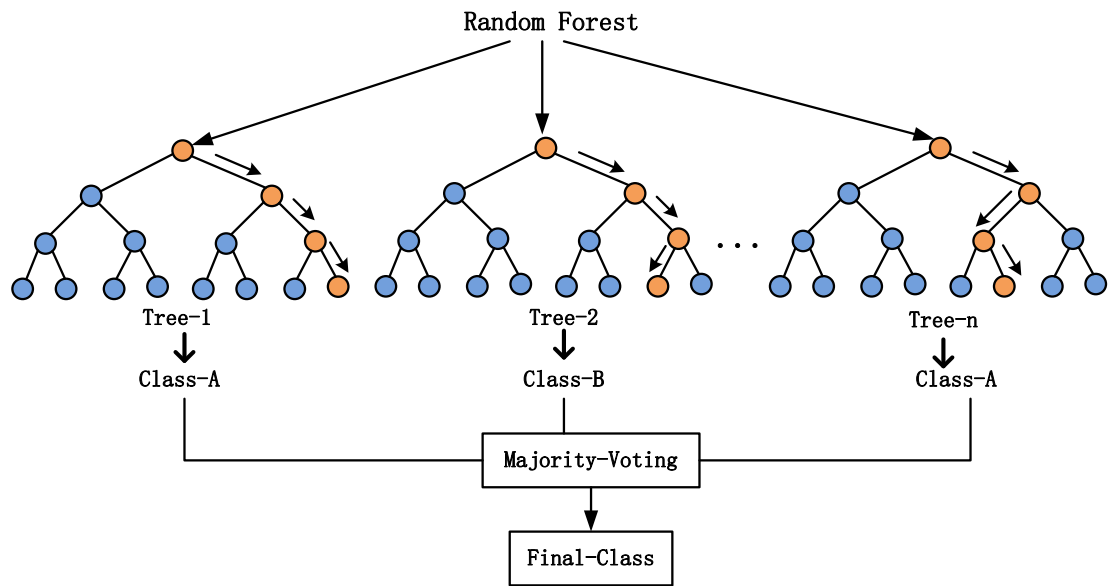


Fig. 2-3 The schematic diagram of random forest [49]

Fig. 2-3 demonstrates the schematic diagram of RF algorithm, and the steps of the algorithm are shown below:

- (1) With RF, n number of random records are taken from the dataset.
- (2) Individual decision trees are constructed for each sample.
- (3) Each decision tree generates an output.
- (4) The final output is considered based on a majority voting for classification.

In summary, the RF is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an

uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

2.5.2 Support vector machine algorithm [50]-[54]

The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space (N- the number of features) that distinctly classifies the data points [50]. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The objective is to find a plane that has the maximum margin, i.e., the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using support vectors, it is possible to maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help build the SVM.

An example of a separable problem in 2-D space is shown in Fig. 2-4. The support vectors marked with the orange circle and the blue square define the margin of largest separation between the two classes. The hyperplane in 2-D space is a line. It can be seen that the margin distance to both support vectors for the hyperplane represented as a solid line is the largest. Another reason for selecting the hyperplane with higher margin is robustness. If we select a hyperplane having low margin, then there is high chance of misclassification [52].

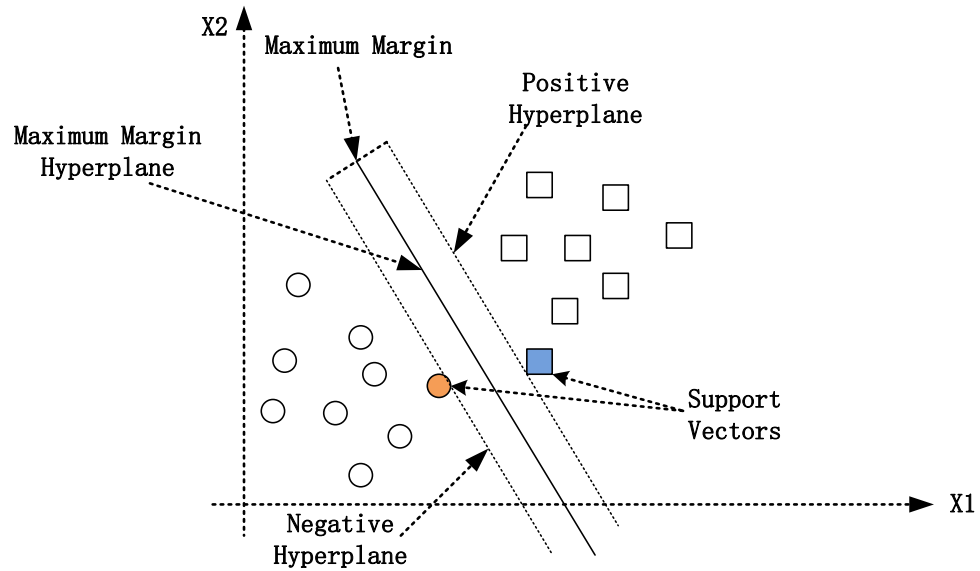


Fig. 2-4 The demonstration of SVM in 2-D space [51]

SVM is basically a linear classifier that classify linearly separable data, but in general, the feature vectors might not be linearly separable. To overcome this issue, the original input space is mapped into a high-dimensional feature space using kernel functions where it becomes linearly separable. Kernel Function generally transforms the input data so that a non-linear decision surface is able to transform to a linear equation in a higher number of dimension spaces. The kernel functions can be different types, such as linear, polynomial, radial basis function (RBF), gaussian and sigmoid, etc. The performance of a SVM classifier is dependent on the choice of a proper kernel function [53]. SVM supports binary classification and separating data points into two classes. For multiclass classification, the same principle is utilized after breaking down the multiclassification problem into multiple binary classification problems. The idea is to map data points to high dimensional space to gain mutual linear separation between every two classes [54].

2.5.3 k-nearest neighbors algorithm [55] [56]

The KNN algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It is easy to implement and understand. The KNN algorithm assumes that similar things exist in close proximity [55]. In other words, similar things are near to each other. KNN works by finding the distances between a query and all the examples in the data. It then selects the specified number of examples (k) closest to the query, and then votes for the most frequent label. The schematic diagram of KNN is shown below [56].

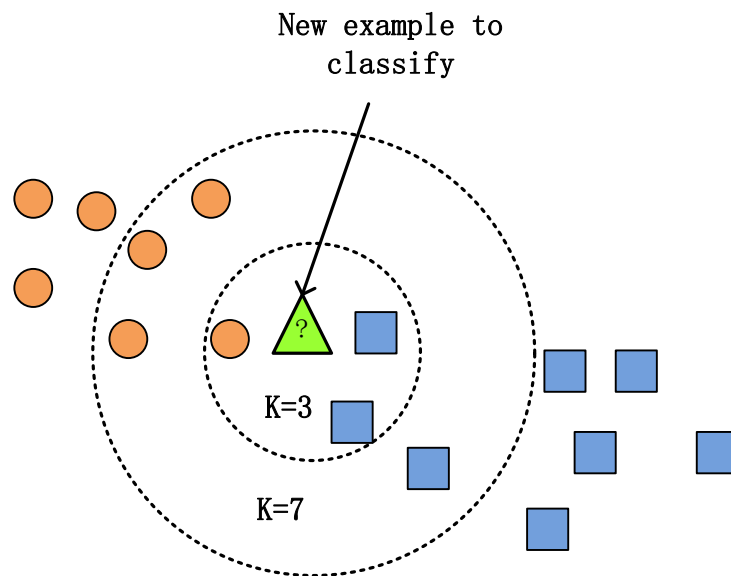


Fig. 2-5 The demonstration of KNN

Fig. 2-5 is an example of KNN classification. The test sample (green triangle) should be classified either to the blue squares or to the orange circles. If $k = 3$, it is assigned to the blue squares because there are 2 squares and only 1 circle inside the inner circle. If $k = 7$, it is assigned to the orange circles (4 circles vs. 3 squares inside the outer circle). To implement a KNN model the following steps are taken [56]:

- (1) Load the data

- (2) Initialize k to the chosen number of neighbors
- (3) For each example in the data, calculate the distance between the query example and the current example from the data. Add the distance and the index of the example to an ordered collection
- (4) Sort the calculated distances in ascending order based on distance values
- (5) Pick the first k entries from the sorted collection
- (6) Get the most frequent class of these entries
- (7) Return the predicted class

2.5.4 Predictor/Feature importance [57]

The importance of the predictors or features used by the ML technique indicates the relative importance of each predictor in estimating the model. For our research, the predictors/features are based on weekly consumption metrics and associated weather patterns. The ML techniques will utilize these features to train its network. In MATLAB, the function of “predictorImportance” computes importance measures of the predictors in a RF tree by summing changes in the node risk due to splits on every predictor, and then dividing the sum by the total number of branch nodes. The change in the node risk is the difference between the risk for the parent node and the total risk for the two children. For example, if a tree splits a parent node (for example, node 1) into two child nodes (for example, nodes 2 and 3), then “predictorImportance” increases the importance of the split predictor by

$$(R_1 - R_2 - R_3) / N_{\text{branch}} \quad (2-7)$$

where R_i is node risk of node i , and N_{branch} is the total number of branch nodes. A node risk is defined as a node error weighted by the node probability:

$$R_i = P_i E_i \quad (2-8)$$

where P_i is the node probability of node i , and E_i is the mean squared error of node i .

2.5.5 Evaluation Metrics [37][58][59]

The key metrics used to evaluate classification algorithms are accuracy, sensitivity, specificity, precision, miss rate, false discovery rate, and false omission rate. All these measurements are derived from the number of true positives, false positives, true negatives, and false negatives obtained when running a set of samples through the supervised-learning classification model. Also, a confusion matrix can be made to display these results. All these main metrics indicates strengths and weaknesses of the classification model which has been built based on the algorithms of RF, SVM and KNN.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (2-9)$$

Sensitivity (TPR - True Positive Rate)

$$\text{TPR} = TP / (TP + FN) \quad (2-10)$$

Miss Rate (FNR - False Negative Rate):

$$\text{FNR} = FN / (FN + TP) \quad (2-11)$$

Precision (PPV - Positive Predictive Value):

$$\text{PPV} = TP / (TP + FP) \quad (2-12)$$

False Discovery Rate (FDR):

$$\text{FDR} = FP / (FP + TP) \quad (2-13)$$

where TP: true positives, FP: false positives, TN: true negatives, FN: false negatives.

CHAPTER 3

3. DATASET 1: RESIDENTIAL CLASSIFICATION BY VALUE, LOT SIZE AND TYPE

3.1 General

Dataset 1 is provided by NB Power and includes:

- ♦ Limited sample data for electricity consumption of 103 residential customers in 15 min intervals between Jan 1st, 2018 to Dec 31st, 2019.
- ♦ Metadata with the following features: lot size, property value, and type of unit.
- ♦ Geographical location of the homes is unknown.

Three metadata features of the NB Power data are shown in Table 3-1: property value, type of unit and lot size. It should be noted that only detached homes are used in this research.

Table 3-1 Dataset 1 of NB Power residences based on available metadata

Group by home/property value		Group by type of unit		Group by lot size	
Value (CAD)	Amount	Type	Amount	Lot size (SQFT)	Amount
1. 0-50k:	5	1. Detached:	85	1. 0-500:	4
2. 50-100k:	23	2. Mobile home:	5	2. 500-1000:	26
3. 100-150k:	28	3. Apartment:	1	3. 1000-1500:	15
4. 150k-200k:	28	4. Attached:	1	4. 1500-2000:	11
5. 200-250k:	3	5. Detached with pool:	8	5. 2000-3000:	7
6. 250k-300k:	6	6. Farm:	3	6. 3000-5000:	13
7. 300k-350k:	4			7. 5000-10,000:	7
8. 350k-400k:	1			8. 10,000-25,000:	10
9. >400k:	5			9. >25,000:	10

In New Brunswick, the summers are warm and humid; the winters are cold and snowy. During cold days, the temperature can drop to -30 °C (-22 °F) or even below, and usually January is the coldest month in New Brunswick. The warmest month in New Brunswick is July with an average maximum temperature of 25°C (77°F). Correspondingly, January and July are initially selected for seasonal data in an attempt to develop residential load

profile clusters given the metadata features given in Table 3-1 and an unknown heating source.

It is noteworthy that the raw data needs to be processed before use. For Dataset 1, there are originally 116 households, but some households do not contain features of lot size, type of unit, or property value leaving only 103 homes to be possibly used in this study. Also, for these 103 homes, their data ranges are different. Some have a data range for the entire two years, while some only cover a couple of months. This issue further reduces the usable data.

Among the normal reading status, another problem is the null reading or zero reading of data for some days. It was assumed that if less than 1/3 day of data is missing, the zero-reading data can be replaced by the average of the nonzero reading data for a few days around the missing day, of the same household, and at the same time. If more than 1/3 day of data reading is zero, the daily data of this household was eliminated. It is understood that processing of missing data in this manner can inadvertently affect the conclusions of the study, but it was deemed avoidable since the data was very limited from the outset.

3.2 Data analysis based on peak-to-mean ratio

After data processing, there are 2621 daily load shapes available in January 2018 and 2889 daily load shapes available in July 2018. It is necessary to combine homes in the following manner to achieve significant a sample space for small and large lots and low-cost and high-cost homes.

1. Small lot size 1-2: 30 samples
2. Large lot size 5-9: 47 samples
3. Low-value homes 1-2: 28 samples
4. High-value homes 5-9: 19 samples

For this basic technique, monthly 2-D scatter data from households based on the metadata of lot size (Fig 3-1 a)) and property value (Fig 3-1 b)) are shown below. The vertical axis is the PMR for each household while the horizontal axis shows the time of occurrence of the peak power for each day.

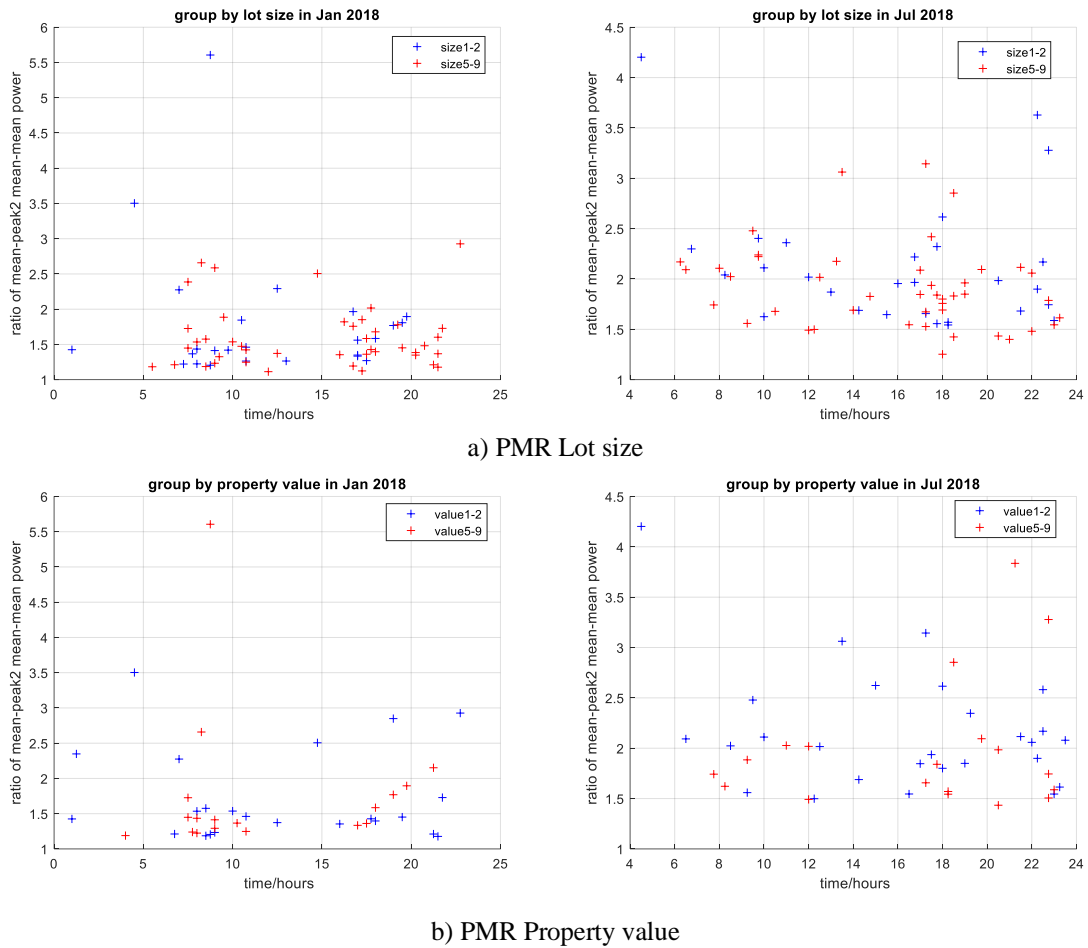


Fig. 3-1 Scatter plots of grouping results based on PMR metric

For Fig. 3-1 a), the PMR is a rather poor metric to separate homes based on lot size. This is somewhat expected since lot size does not always correlate to home size and heating type is unknown. Similarly, for Fig. 3-1 b), PMR results based on home/property value are given. The same conclusion can be drawn as with lot size, where PMR cannot be used to consistently group homes based on their value. Numerous characteristics can affect

residential consumption with key metadata features such as heating type and home size being a significant contributor. Both of these are unknown values for Dataset 1.

In summary, the above results show no clear boundaries for residential groupings based on a single metadata feature of lot size or property value. The same result is similar for all calendar months since these metadata features individually do not provide enough information for customer discrimination irrespective of the time of year.

3.3 Grouping data based on combined metadata features

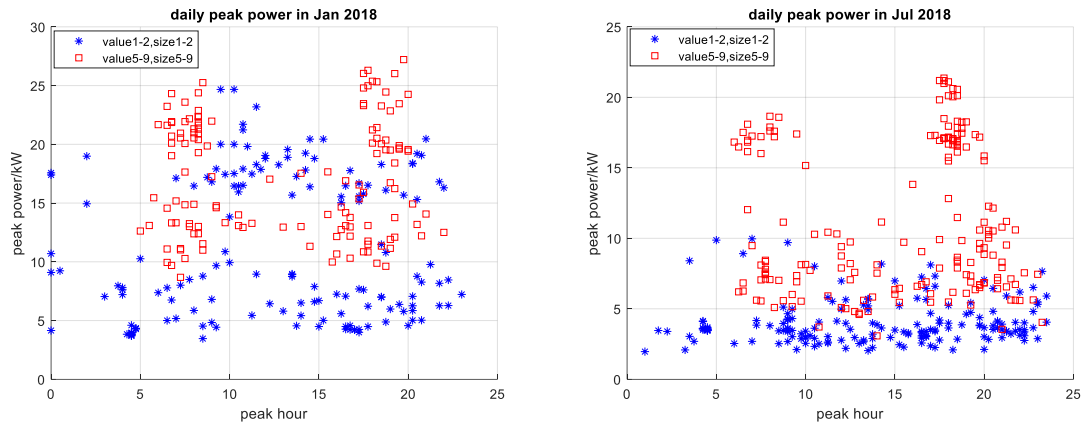
This section provides an assessment of whether combining metadata features can improve home consumption separability. From Table 3-1, it is clear that limited features can be chosen for the combination. It seems logical to combine lot size and property given the assumption that a small lot size with low property value can be considered as a small house in terms of square footage, while in comparison, a large lot size with high property value could be to be a large house. In this way, it may be possible to separate load profiles between the groups of low property value/small lot versus high property value/large lot.

3.3.1 The assessment of combining metadata features

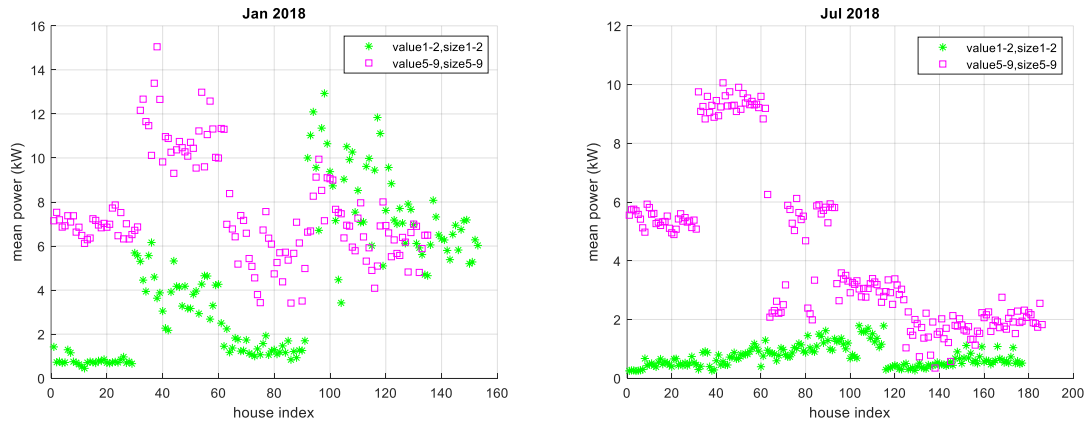
Two groups are selected by combining metadata features of property value and lot size. Group 1 (property value 1-2 and lot size 1-2) has low price homes on small lots, while Group 2 (the property value 5-9 and lot size 5-9) is with high price homes on large lots. This was done to hopefully create a larger separation of load profiles with the assumption that Group 1 and 2 represent small and large sized homes, respectively.

For Fig 3-2 a), it is generally found that, Group 1 has lower peak power than Group 2, while in January (cold days), there is quite a bit of overlap. In July (hot days), the separability of Group 1 and Group 2 is much better than in January. The house index in Fig

3-2 b) indicates the various homes represented in Group 1 and Group 2, and similar results can be obtained from simple mean power plots. PMR results also show similar results. There is some overlap of Group 1 and Group 2 in January based on mean power while the boundary of mean power distribution of Group 1 and Group 2 is much clearer in July than in January.



a) Monthly peak power and peak hour



b) Daily mean power for each home

Fig. 3-2 Peak power and mean power distribution for combining features

The results indicate there is some merit in assuming that lower value and smaller lot size homes have lower power consumption, while higher value and larger lot size homes have higher power consumption. On cold days, differences are less clear given the unknown nature of the heating source, efficiency of the home and absolute certainty of the

home square footage. This conclusion corresponds to Chapter 1 which mentioned that in New Brunswick, the average household energy was comprised of approximately 48 percent electric space heating, and 2 percent air conditioning. Electric space heating occupies a major part of residential consumption during cold days, while unknown heating source type in Dataset 1 makes the load profile grouping in cold days difficult.

3.3.2 Representative daily load shapes based on K-means clustering

For this method, again January and July 2018 are used as cold days and hot days, respectively. The K-means clustering algorithm is applied to cluster the dataset and get the representative daily load shapes. The load shapes represented by the centroids are helpful for shape analysis such as T-test analysis. Both the load patterns in January and July are grouped into 15 clusters and every load shape in January and July is mapped to the closest shape cluster, respectively. Fig. 3-3 a) shows the representative load shapes which are the centroids of K-means clustering in January and July, respectively. Resulting cluster centroids are based on the daily load patterns of each user with $24 \times 4 = 96$ dimensions and are used to form the shape dictionary.

The input data is again partitioned into two groups based on the combining metadata features as mentioned before, Group 1 to be considered low property value/small lot and Group 2 to be considered high property value/large lot. Daily load patterns for each group are shown in Figure 3.3 a). Every load pattern of Group 1 in January is assigned to one of the 15 resulting cluster centers for the month of January that minimizes the squared error. The number of load shapes of Group 1 mapped to each cluster centroid is counted. The same operation applies to load patterns of Group 2 to get the load shapes distribution. All the steps for Group 1 and Group 2 in January are repeated in July. Plots in Fig 3-3 b)

demonstrate the number of load shapes in each cluster and ideally, two single peaks each representing a different group shape at different cluster indices would be ideal. However, these results are somewhat inconclusive given numerous overlapping peaks at different indices for each group and month. That said, there are still some unique cluster centroids, especially for the month of July.

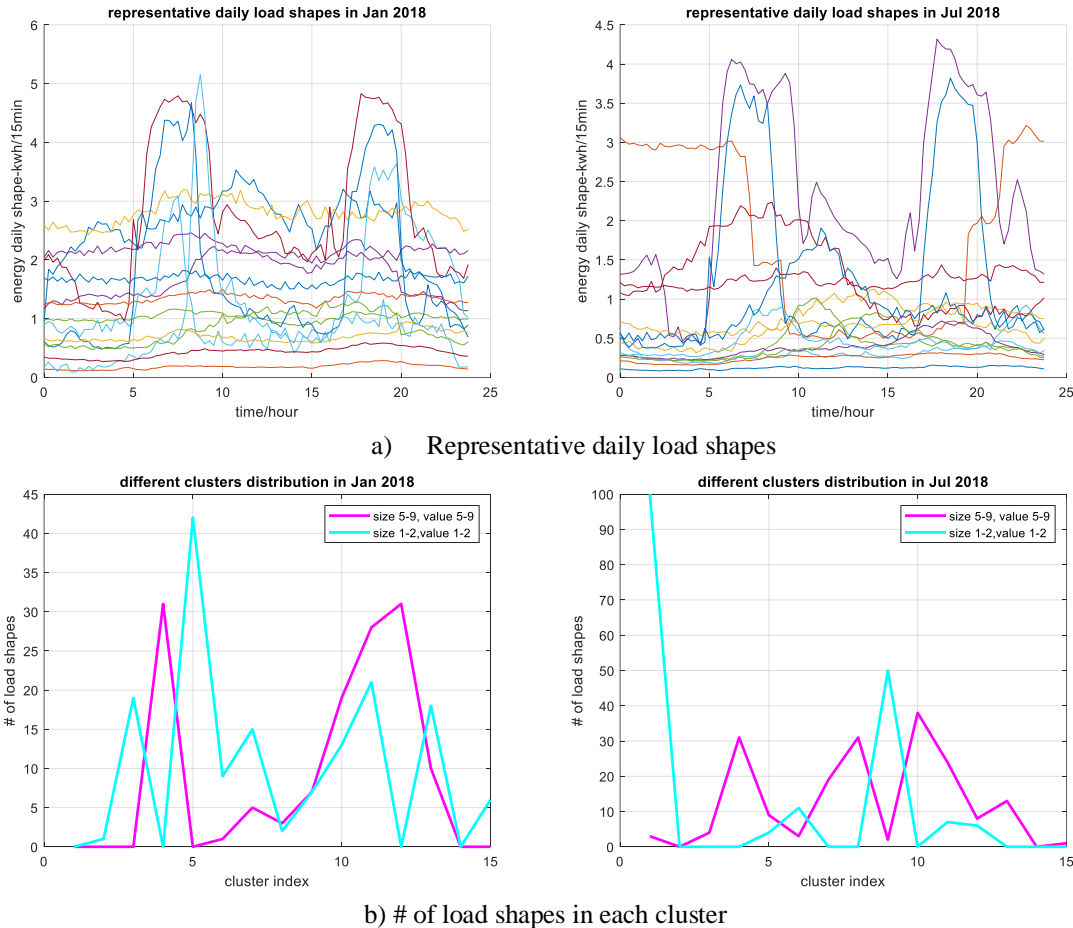


Fig. 3-3 Two feature metadata grouping results based on K-means clustering

This is in agreement with the previous results where for cold days, the unknown heating source can make it difficult to distinguish the daily consumption between Group 1 and Group 2, so there are many more chances the load shapes of these two groups contribute to the same cluster. In warmer days, it is not common for the residents of New Brunswick to have a cooling system, so the daily consumption of these two groups mainly

may follow the assumption that lower cost homes have low power consumption than more expensive homes and this contributes to separable clusters in the July results.

3.3.3 Student's T-test

Again, a comparison is made of Group 1 and Group 2. The frequency of each representative load shape is compared in two groups. Since the frequency is an estimate of the true frequencies of load shapes, a two sample T-test is utilized (Eq. 2-4). The T-test is used to assess the statistical significance of load profiles differing in shape for the two groups used above. A two tailed test result is plotted in Fig. 3-4 (assuming 5% significance level, split 2.5% each on either side). The blue line indicating T-distribution value is selected by significance level and d.f. (Eq. 2-5) according to Appendix A. The red dots between the two blue lines indicate common representative load shapes in both Group 1 and 2, that is, no statistical difference exists (i.e. the null hypothesis). The red dots above the upper and lower blue line mean more frequent representative load shapes in Group 1 and Group 2, respectively. Correspondingly, the plots show that in January there are 5 more frequent representative load shapes in Group 1, 2 more frequent representative load shapes in Group 2, and 6 common representative load shapes in both groups. In July, there are 3 more frequent load shapes in Group 1, 7 more frequent load shapes in Group 2, and only 3 common load shapes in both groups.

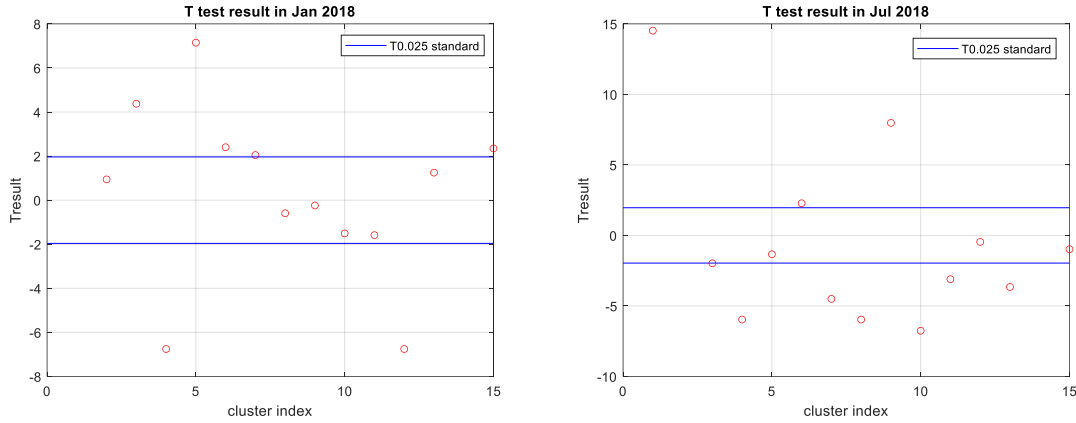


Fig. 3-4 Two feature metadata grouping T-test results

Table 3-2 T-test results 2018

T-test result in 2018	Jan	Jul
$P(C_i \text{Group 1}) > P(C_i \text{Group 2})$	5	3
$P(C_i \text{Group 1}) < P(C_i \text{Group 2})$	2	7
$P(C_i \text{Group 1}) = P(C_i \text{Group 2})$	6	3
Total	13	13

Table 3-2 show T-test probability results as described in section 2.4.2. Here, equivalent probabilities indicate a poor separability between groups which in January the two groups have a **53.85%** (7/13) total probability of separating load profiles between Group 1 and 2, which is not much better than chance. On the other hand, for July, there is **76.92%** (10/13) probability of load shapes being dissimilar, which is a fair indicator of different load shapes being specifically associated with each of these two groups. It should be noted that these results cannot be taken with a high level of confidence given the limited sample space and resulting large standard error.

3.3.4 Monthly T-test analysis

The above statistical method is expanded to the whole year of 2018. For each month, representative daily load shapes are obtained by K-means clustering. The load patterns of two groups in each month are assigned to their corresponding closest centroids,

respectively. Finally, T-test results are provided to test the statistical difference between the two groups.

Table 3-3 demonstrates the T-test result of the two groups for each month in 2018 and shows that Group 1 and 2 load separability with combined metadata performs well in hot days such as June and July, where the difference of the two groups can be high up to **75%** or more. In mild days such as May and September, the two groups of daily shapes can reach at least a 60% difference, which is acceptable. During cold days, such as January and February, the difference of the two groups is difficult to distinguish, with results under 60% but more than 50%. Therefore, the ability of T-test results to reveal statistical differences in residential load profile is limited to warmer months and difficulties lie for all other times of the year.

Table 3-3 T-test results in 2018

Month	# of available daily shapes	T test Group 1 vs Group 2			
		Combined greater than/less than probabilities	Equivalent probability	Percent separability	Percent inseparability
Jan-18	2621	7	6	53.85%	46.15%
Feb-18	2472	7	6	53.85%	46.15%
Mar-18	2815	7	6	53.85%	46.15%
Apr-18	2727	8	5	61.54%	38.46%
May-18	2810	9	4	69.23%	30.67%
Jun-18	2822	9	3	75.00%	25.00%
Jul-18	2889	10	3	76.92%	23.08%
Aug-18	2908	9	5	64.29%	35.61%
Sep-18	2688	9	6	60.00%	40.00%
Oct-18	2890	7	6	61.54%	38.46%
Nov-18	2758	7	5	58.33%	41.67%
Dec-18	2844	8	6	57.14%	42.86%

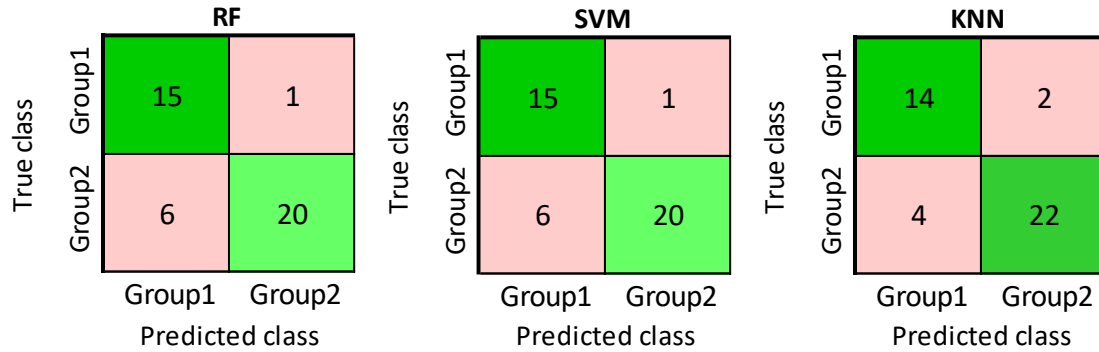
3.4 Classification based on ML algorithms

Although statistical analysis yielded inconsistent results, ML techniques, as described in section 2.5, are tested on this dataset. RF, SVM and KNN are applied to Dataset 1 for the purposes of assessing the separability of Group 1 and Group 2 in both summer and winter months, with a specific focus on the winter months, where statistical analysis fails to provide a differentiation between these two groups.

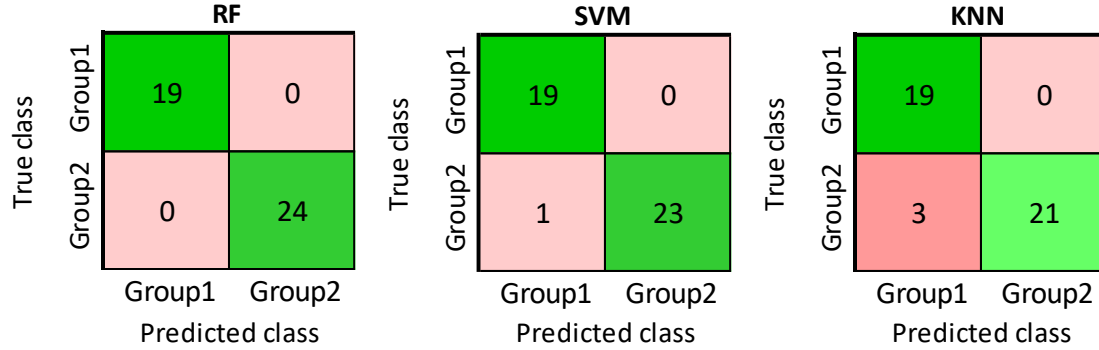
For the supervised ML algorithms, winter consumption data in the months of December and January is selected for training, and the consumption data in February is used for testing. Similarly, for summer, consumption data in the months of June and July is selected for training, and the consumption data in August is used for testing. Both training and testing data were attempted at least 5 times to get representative classification results. MATLAB is used extensively for the ML algorithms, along with RStudio to extract the features of the consumption data. Feature vectors were derived from consumption data collected at 15-min granularity over one week (7 days). The package of “SmartMeterAnalytics” [36] in RStudio is used to extract the 62 consumption features (shown in Appendix B) used for the classification of the ML model, and the description of each feature is listed in [35]. No weather information was used since no metadata exists indicating residential location.

For these three algorithms, the predictors are the 62 consumption features, and the responses are Group 1 or Group 2. To use the algorithm functions in MATLAB, several functional parameters were evaluated and the best results are given. For RF, as illustrated in Fig.2-3, the number of trees is set as 50 ($n=50$) to get relatively high accuracy and less time cost, and for SVM as demonstrated in Fig. 2-4, the Gaussian kernel function is chosen

to map the input data into higher dimensional space to make it linearly separable, and for KNN as demonstrated in Fig. 2-5, k is set to 10 yielded the best results.



a) Winter month results using 62 consumption features



b) Summer month results using 62 consumption features

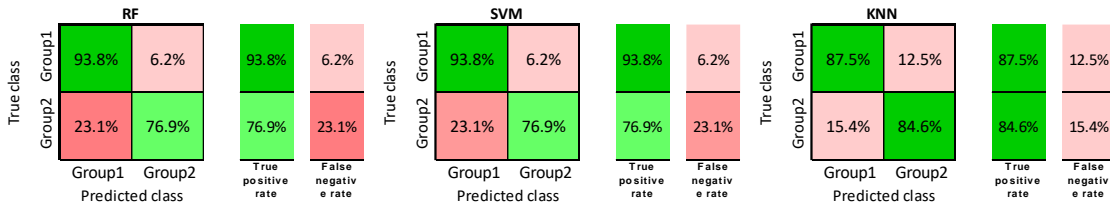
Fig. 3-5 Supervised learning load classification results

Fig. 3-5 shows a typical classification result of both winter and summer months with the three different ML algorithms, respectively. The green color in the confusion matrix indicates the number of load shapes classified correctly, while the red color means the number of load shapes that have been categorized to the wrong class. Both winter and summer data get very good classification results, which differs from those based Student's T-test which performed poorly for winter months.

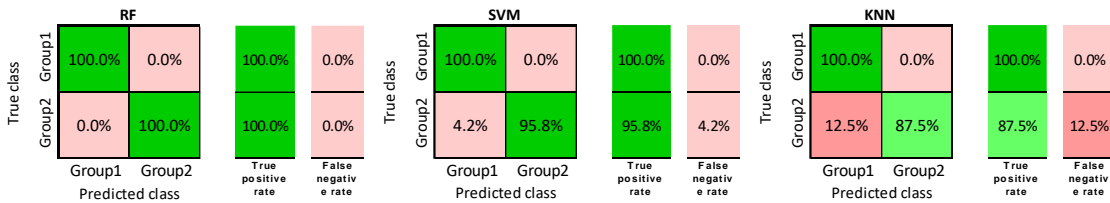
Table 3-4 Consumption data classification results

Groups	Algorithms	Group1	Group2	Subtotal	Accuracy
Winter data (62 features)	RF	15/16	20/26	35/42	83.3%
	SVM	15/16	20/26	35/42	83.3%
	KNN	14/16	22/26	38/42	85.7%
Summer data (62 features)	RF	19/19	24/24	43/43	100.0%
	SVM	19/19	23/24	42/43	97.7%
	KNN	19/19	21/24	40/43	93.0%

The accuracy of the classification results calculated using Eq. (2-9), is shown in Table 3-4. It is seen that both winter and summer months get high classification accuracy results. This is especially interesting since it shows the ability of the three ML techniques to discern Group 1 and 2 load profile differences in winter months, which the previous student's T-test could not.



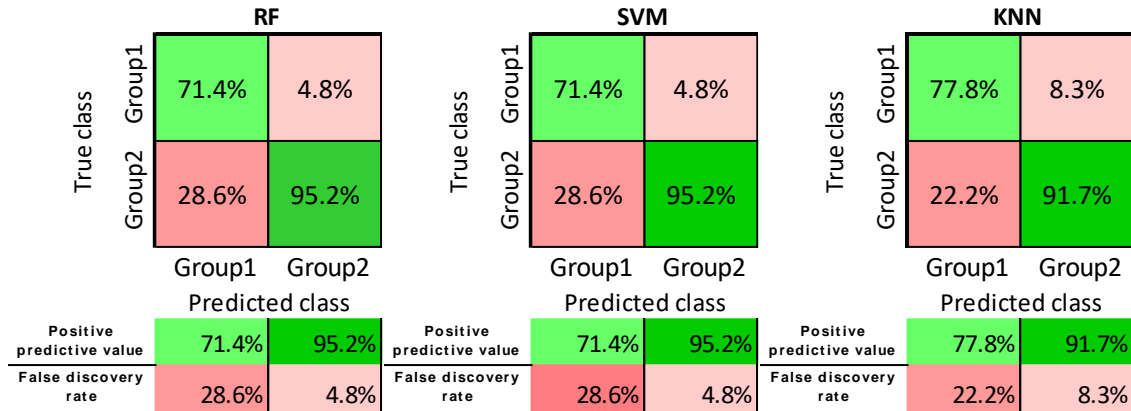
a) Winter data results using 62 consumption features



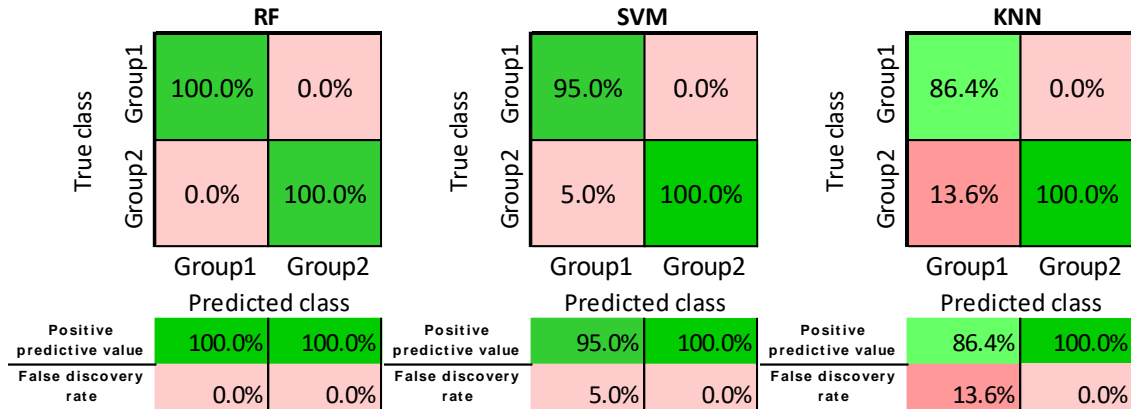
b) Summer data results using 62 consumption features

Fig. 3-6 True positive rate and false negative rate of different subset

Fig. 3-6 shows the True Positive Rate (TPR), as found using Eq. (2-10), and False Negative Rate (FNR), as found using Eq. (2-11), of the two subsets with the three different supervised ML algorithms, respectively. Generally, Group 1 has a higher TPR than Group 2 in both winter and summer months, while summer month performs 100% of TPR in Group 1.



a) Winter data results using 62 consumption features



b) Summer data results using 62 consumption features

Fig. 3-7 Positive predictive value and false discovery rate of different subset

Fig. 3-7 shows the Positive Predictive Value (PPV), as found using Eq. (2-12), and False Discovery Rate (FDR), as found using Eq. (2-13), of the three ML models, respectively. In general, Group 2 has a higher PPV than Group 1 in both winter and summer months, while summer month performs 100% of PPV in Group 2.

Given the limitations of the available consumption data for Dataset 1, these ML results will not be consistent with more training, testing data and an increased number of classification groups. The binary grouping of this test case combined with the rather limited sample space and variability will accentuate the ML classification accuracy for this test case.

3.5 Analysis and summary

In this chapter, Dataset 1 with the combined metadata features of lot size and property value were used for load classification and analysis. Simple consumption metrics along with K-means algorithm and Student's T-test had some success for summer months, but colder winter months proved difficult. Supervised ML algorithms of RF, SVM and KNN had better success in both summer and winter month categories.

Firstly, the PMR metric is used to perform preliminary results based on single feature metadata lot size and property value. The results show no clear boundaries for residential groupings based on a single metadata feature of lot size or property value. The same result is expected for all calendar days since these features individually do not provide enough information for customer discrimination irrespective of the time of year.

Then combining metadata features of lot size and property value is introduced. K-means clustering, and Student's T-test are used to implement the load shape analysis. The results show potential in combining features of lot size and property value to provide insight into load profiles given a residential home fitting in the category of Group 1 or 2. This study is not 100 per cent conclusive but statistical evidence indicates there is some difference in consumption patterns for Group 1 versus Group 2, especially for the summer months. Limited success was found for winter months.

Finally, ML techniques are tested on this dataset. Given its limited sample size, the results obtained by applying these models should be considered exploratory in nature, and the significance of which should be tempered. RF, SVM and KNN are applied to Dataset 1 for the purposes of assessing the separability of Group 1 and Group 2 in both summer and winter months. The results show that summer months classification has the better classification result which corresponds to the in the T-test analysis. However, both winter and summer months get high accuracy results, which shows the ML algorithms ability to extract load profile information in winter months, where the student's T-test failed to do so.

CHAPTER 4

4 DATASET 2: RESIDENTIAL HEAT SOURCE CLASSIFICATION

In Chapter 1, it was shown that electric space heating occupies 48% of energy consumption. As expected, the type residential heating source is a very important factor in determining load profiles of individual consumers and aggregations of a large number of consumers. HPs are modern systems generally used to heat and cool rooms by using electricity to convert natural energy from ground water, the earth or air into usable heat energy. Grid operators can benefit from greater deployments of HPs to improve energy efficiency in colder and hotter months of the year, as well as improve the overall load profile of the grid. For this part of the thesis, the ability to detect various heating types is especially important when electric BB heating is the dominant type, as in the case of New Brunswick. This part of the thesis explores the goal of using statistical and ML techniques to detect the presence of HPs from only load consumption patterns. This is an important determination in assessing future load profiles for a particular home.

4.1 General

Dataset 2 is supplied by NB Power and is from homes in the area of Shediac, NB. This geographical metadata will be used for ML techniques where historical weather data is needed. Dataset 2 provides residential consumption data from the year 2020 with only the metadata of heating source type. After the same data preprocessing as with Dataset 1, the available homes and associated heating type are shown.

Table 4-1 Dataset 2 of NB Power residences based on heating source

Dataset 2 (504)		
Primary heating source		# of users
non-electric (27)	oil	13
	propane	2
	wood	12
electric (477)	central boiler	15
	heat pump air (HP)	81
	heat pump geo	6
	baseboard (BB)	114
	radian	8
	mini-split heat pump (MS)	253

An initial study on this dataset will use statistical analysis and K-means clustering to assess load profile differences between electric and non-electric (NE) heating sources. More importantly, electric heating sources are further classified into HP, BB, and MS categories using statistical and ML techniques. For the latter, open-source weather data will also be used along with the consumption data. MATLAB is used extensively for the ML algorithms, along with RStudio to extract the features (shown in Appendix C) of the consumption and weather data used for the training of the ML model.

4.2 Grouping data by heating source

4.2.1 Heating source analysis

After data processing, the load consumption for the whole year of 2020 is chosen. As was done in the previous chapter, the usual typical cold days (January) and hot days (July) of the residential load dataset are compared. There are 15565 daily load shapes available in January 2020 and 15598 daily load shapes available in July 2020.

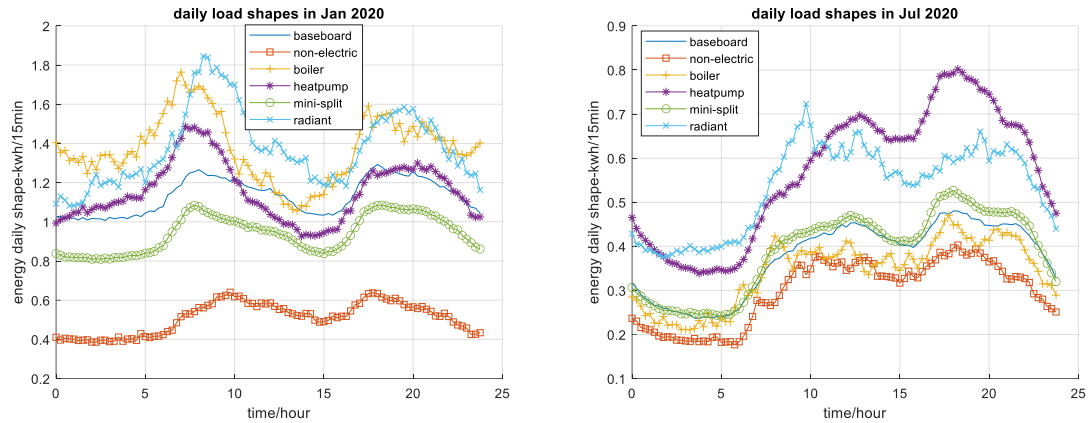


Fig. 4-1 Average daily load shapes for different heating sources

Fig. 4-1 shows the average daily load shapes of different heating sources in January and July. Certainly, the NE heating homes have the lowest energy consumption both in January and July, while the average energy usage is a little bit higher in January than in July. BB heating is commonly used in New Brunswick but cannot be used for cooling. In July, BB power consumption is slightly higher than NE heating which can be due to a considerable proportion of BB homes having available MS units for cooling in the summer. This explains why MS heating homes have a similar load profile to, but slightly lower, than BB.

Like a MS unit, a HP can be used both in heating and cooling, these units are very efficient, using approximately 1/2 to 1/3 of energy compared to BB [60]. The problem with HP installations is the internal ductwork needed to distribute the heat throughout the home. As a consequence, newer modern homes will be the primary locations of HP installations since retrofitting older homes with ductwork is quite costly. MS units, however, are popular since they can be installed for localized heating and cooling and internal ductwork is not needed. In any event, HP and MS heating should utilize less energy for similar sized homes when compared to BB heated homes but is less conclusive for home of varying sizes.

Unfortunately, home size is not available as a metadata feature for Dataset 2, so simply examining residential energy usage using simple metrics like PMR, peak consumption, average consumption etc. will usually not be a good indicator of heating type. This is illustrated by the energy usage for four key sources is given below in Table 4-2:

Table 4-2 Dataset 2: Average energy consumption in each month

Month	# of daily shapes	# of nonzero daily shapes	Mean energy kWh/15min				
			BB	NE	HP	MS	Average
Jan-20	16062	15565	1.1317	0.5103	1.1560	0.9393	1.0239
Feb-20	15210	14703	1.0844	0.4927	1.0936	0.9093	0.9850
Mar-20	16202	15692	0.9571	0.4671	0.9294	0.8048	0.8639
Apr-20	15717	15210	0.7332	0.4128	0.7068	0.6333	0.6715
May-20	16211	15703	0.5155	0.3348	0.5643	0.4683	0.4961
Jun-20	15674	15131	0.3884	0.2954	0.5510	0.3933	0.4153
Jul-20	16108	15598	0.3754	0.2983	0.5682	0.3897	0.4125
Aug-20	15542	14544	0.3681	0.2896	0.5691	0.3865	0.4084
Sep-20	15531	15028	0.3537	0.2626	0.4817	0.3639	0.3776
Oct-20	15984	15477	0.4551	0.3201	0.5162	0.4369	0.4554
Nov-20	15299	14759	0.6452	0.3939	0.6671	0.5764	0.6100
Dec-20	15369	14719	0.8533	0.4745	0.9102	0.7480	0.8000

BB and MS usage covers about 75% of heating in New Brunswick. HP usage is about 15% which indicates some newer homes present in Dataset 2.

Since only a single metadata feature along with general geographical location is provided for Dataset 2, it is only possible to discriminate load profiles by heating source. The commonly used electric space heating sources (BB, HP and MS) are initially chosen to compare against with NE space heating, respectively. This is a relatively simplistic evaluation and was only done to ensure that electric and non-electric heating sources can be distinguished with statistical analysis.

4.2.2 Load shape analysis: T-test results

For the sake of brevity only the statistical T-test results are provided. The load profile analysis leading up to this is exactly the same as given in Chapter 3. Three T-test results are provided: BB vs NE, HP vs NE and MS vs NE. Although T-tests are fairly rudimentary, they can clarify distinct differences between electric and non-electric sources depending on the time of year.

Table 4-3 demonstrates the T-test results of the two groups of BB and NE for January and July of 2020. The T-test results show that the data distinguished by heating source performs very well on cold days, such as January, where the difference of the two groups can be high up to almost 80%. During hot days such as July, the difference of the two groups is very difficult to distinguish with only about 20% or under, which means almost 80% of the load shapes are similar.

Table 4-3 T-test results in 2020

Month	# of daily shapes	# of available daily shapes	T test (BB vs NE)			
			Combined greater than/ less than probabilities	Equivalent probability	Percent separability	Percent inseparability
Jan-20	16062	15565	35	9	79.55%	20.45%
Jul-20	16108	15598	10	50	16.67%	83.33%

As expected, the T-test results show a significant difference between BB heating and NE space heating load shapes on cold days. BB space heating has limited impact on hot days, so there is no statistical difference of these two sources in summer as energy usage of air-conditioning is considered limited in New Brunswick.

For the second preliminary test, homes with HPs are compared against NE heated homes. The T-test is used to assess the statistical significance of load profiles differing in shape for the two sources used above and Table 4-4 show that in January the two

comparisons have a **71.93%** (41/57) probability of load shapes difference which is a fair indicator of different load shapes being representative of these two sources. For July, there is a **64.41%** (38/59) probability of indicating different load shapes. Both cold and hot daily load shapes have a clearer boundary since HPs can be used for both heating and cooling and thus have impact on both cold days and hot days

Table 4-4 T-test result (HP vs NE)

Month	# of daily shapes	# of available daily shapes	T test (HP vs NE)			
			Combined greater than/less than probabilities	Equivalent probability	Percent separability	Percent inseparability
Jan-20	16062	15565	41	16	71.93%	28.07%
Jul-20	16108	15598	38	21	64.41%	35.59%

For the T-test is used to assess the statistical significance of load profiles differing in shape for MS and NE heating. Table 4-5 show that in January the two sources have a **62.75%** (32/51) probability of load shapes being dissimilar which is a fair indicator of different load shapes being represented in these two sources. For July, however, there is only a 26.67% (16/60) probability of indicating different load shapes, other words, there is a **73.33%** (44/60) probability of similar load shapes. There is clearly more difficulty in differentiating MS from NE in the summer months. This could be due the smaller British thermal unit (BTU) sizing of MSs compared to larger HP units which are generally installed for centralized cooling and heating.

Table 4-5 T-test result (MS vs NE heating)

Month	# of daily shapes	# of available daily shapes	T-test (MS vs NE)			
			Combined greater than/less than probabilities	Equivalent probability	Percent separability	Percent inseparability
Jan-20	16062	15565	32	19	62.75%	37.25%
Jul-20	16108	15598	16	44	26.67%	73.33%

4.3 Load profile variations of electric space heating types

As stated previously, BB, MS and HPs constitute about 90% of the space heating in New Brunswick. Understanding the characteristics of these three different heating types will be helpful for load assessment. Once more, K-means and Student's T-test are used to analyze the separability of load shapes among these three electric heating types, and limitations of this process are provided.

The K-means clustering algorithm is applied to cluster the dataset and get the representative daily load shapes. Both the load patterns in January and July are grouped into 60 clusters, and every load shape in January and July is mapped to the closest shape cluster, respectively. Fig. 4-2 shows the representative load shapes which are the centroids of K-means clustering in January and July, respectively. It is obvious that residential households consume more energy on cold days than on hot days in New Brunswick. Spacing heating contributes to the majority of energy differences between cold days and hot days.

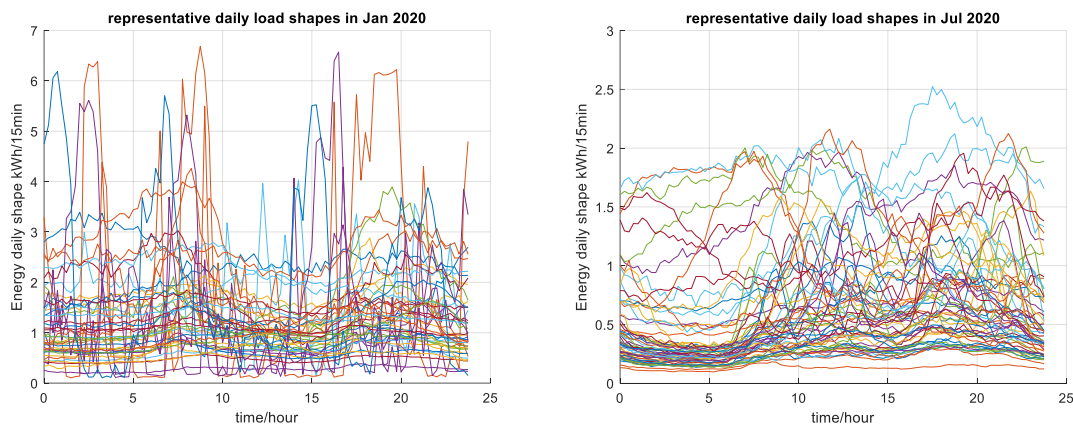
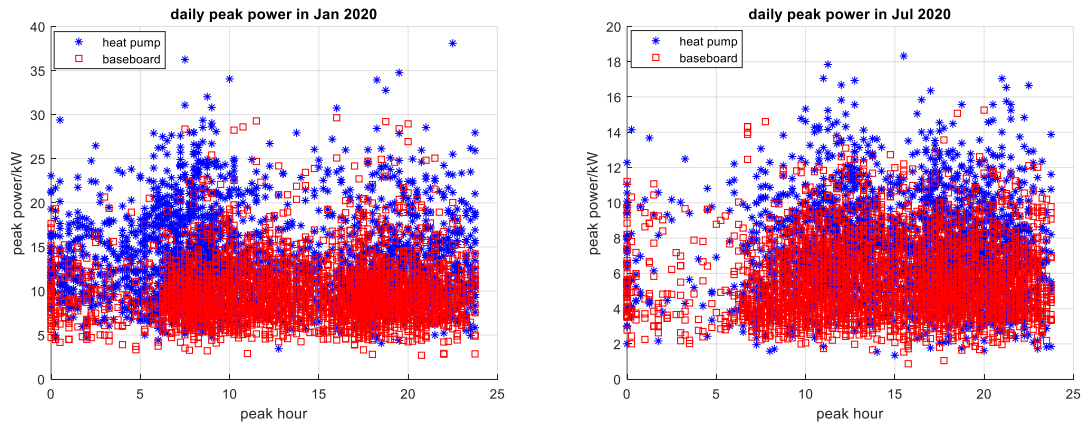


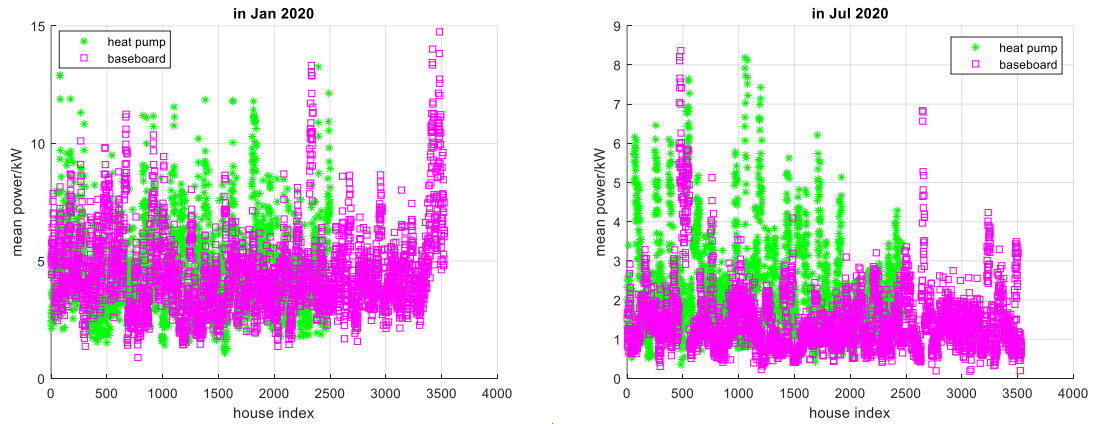
Fig. 4-2 Representative daily load shapes

4.3.1 HP vs BB

Two heating sources are selected from Dataset 2: HP homes vs BB homes. Fig. 4-3 shows the boundary of peak power is clearer in cold days than in hot days. This figure shows some separability using mean power in July but less so for January.



a) Peak power and peak hour



b) Daily mean power for each daily shape

Fig. 4-3 Peak power and mean power (HP vs BB)

All the available load patterns in January and July are grouped into 60 clusters (Fig. 4-2) based on the K-means algorithm, respectively. Plots in Fig 4-4 demonstrate the number of load shapes in each cluster and show much more distinguishing characteristics in the cluster index for July when compared with January between HP and BB.

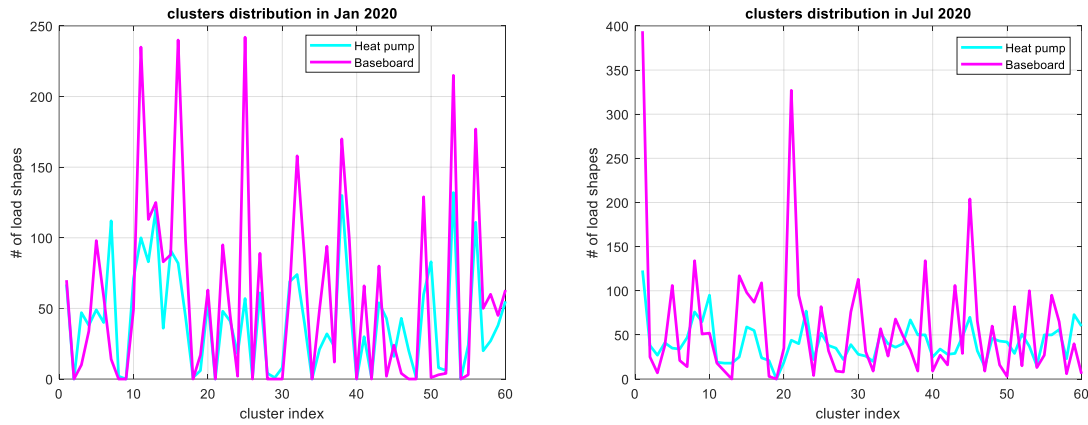


Fig. 4-4 # of load shapes in each cluster (HP vs BB)

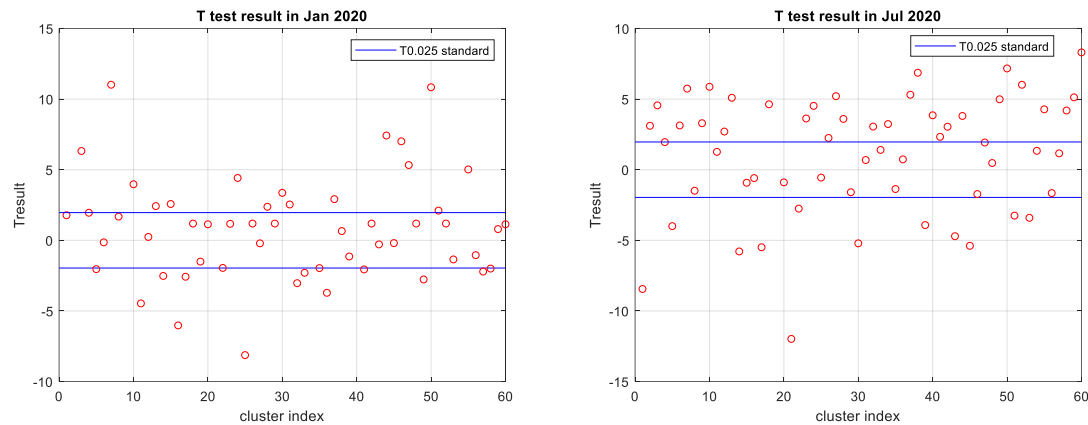


Fig. 4-5 T-test results (HP vs BB)

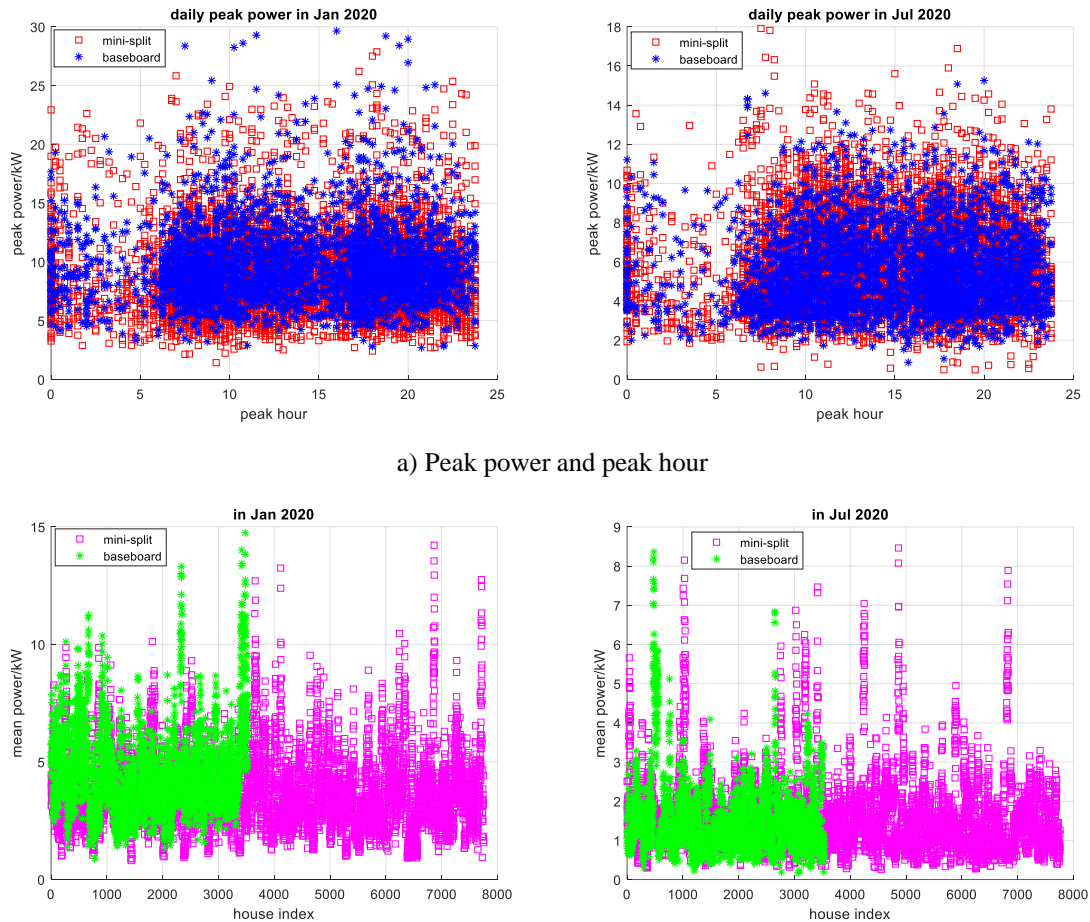
The T-test is used to assess the statistical significance of load profiles differing in shape for the two heating sources used above. Fig. 4-5 and Table 4-6 show that in July the two sources had a **69.49%** (41/59) probability of load shapes difference which is a fair indicator of different load shapes being represented for these two sources. For January, there is only a **55.56%** (30/54) probability of indicating different load shapes. Even though HPs have average higher consumption in winter, their differences are not enough to distinguish these two sources clearly. In summer, the impact of HP cooling is very obvious, and there is better separability of load shapes.

Table 4-6 T-test result (HP vs BB)

Month	# of daily shapes	# of available daily shapes	T-test (HP vs BB)			
			Combined greater than/less than probabilities	Equivalent probability	Percent separability	Percent inseparability
Jan-20	16062	15565	30	24	55.56%	44.44%
Jul-20	16108	15598	41	18	69.49%	30.51%

4.3.2 MS vs BB

For this case, a comparison is made between MS vs BB homes. Fig. 4-6 shows a limited boundary using the metrics of daily peak power and daily mean power for the month of January, but almost no difference between these two sources in July.



a) Peak power and peak hour

b) Daily mean power for each daily shape

Fig. 4-6 Peak power and mean power (MS vs BB)

All the available load patterns in January and July are grouped into 60 clusters (Fig. 4-2) based on the K-means algorithm, respectively. Plots in Fig. 4-7 demonstrate the number of load shapes in each cluster and show that distinguishing distribution features as function of cluster index for January and July is not obvious.

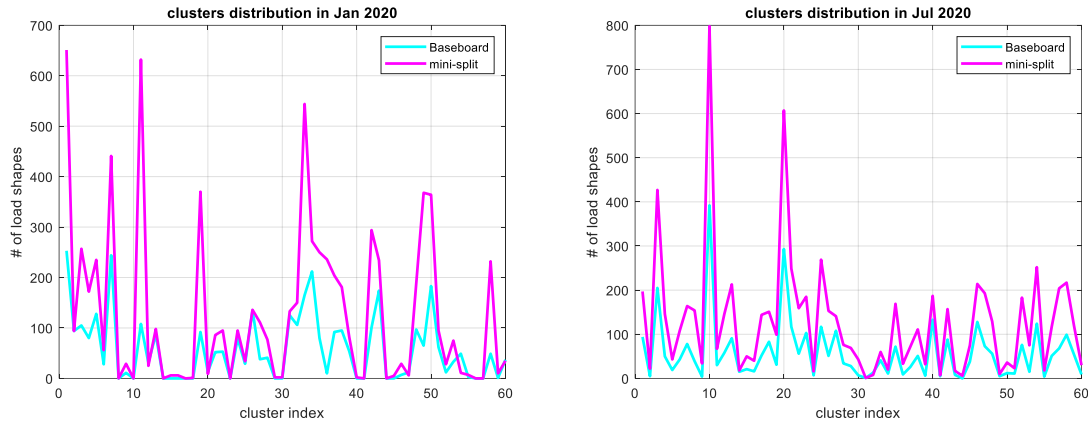


Fig. 4-7 # of load shapes in each cluster

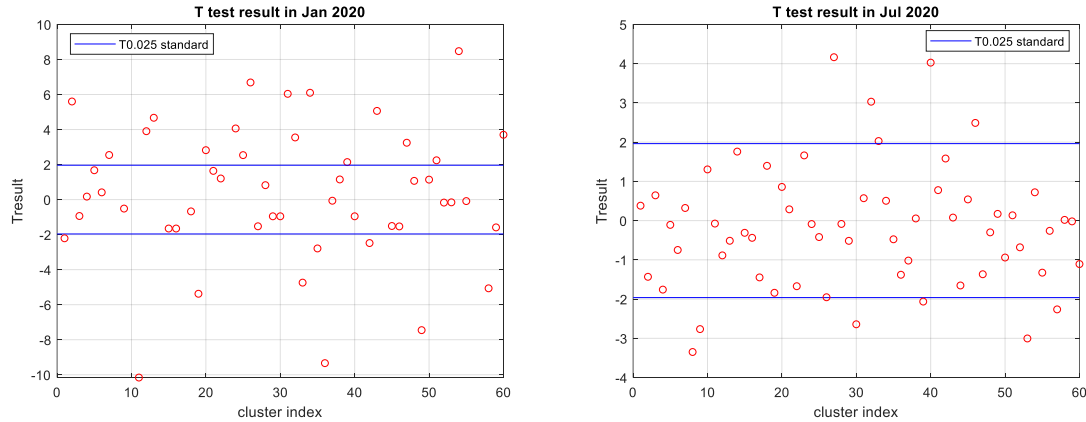


Fig. 4-8 T-test results (MS vs BB)

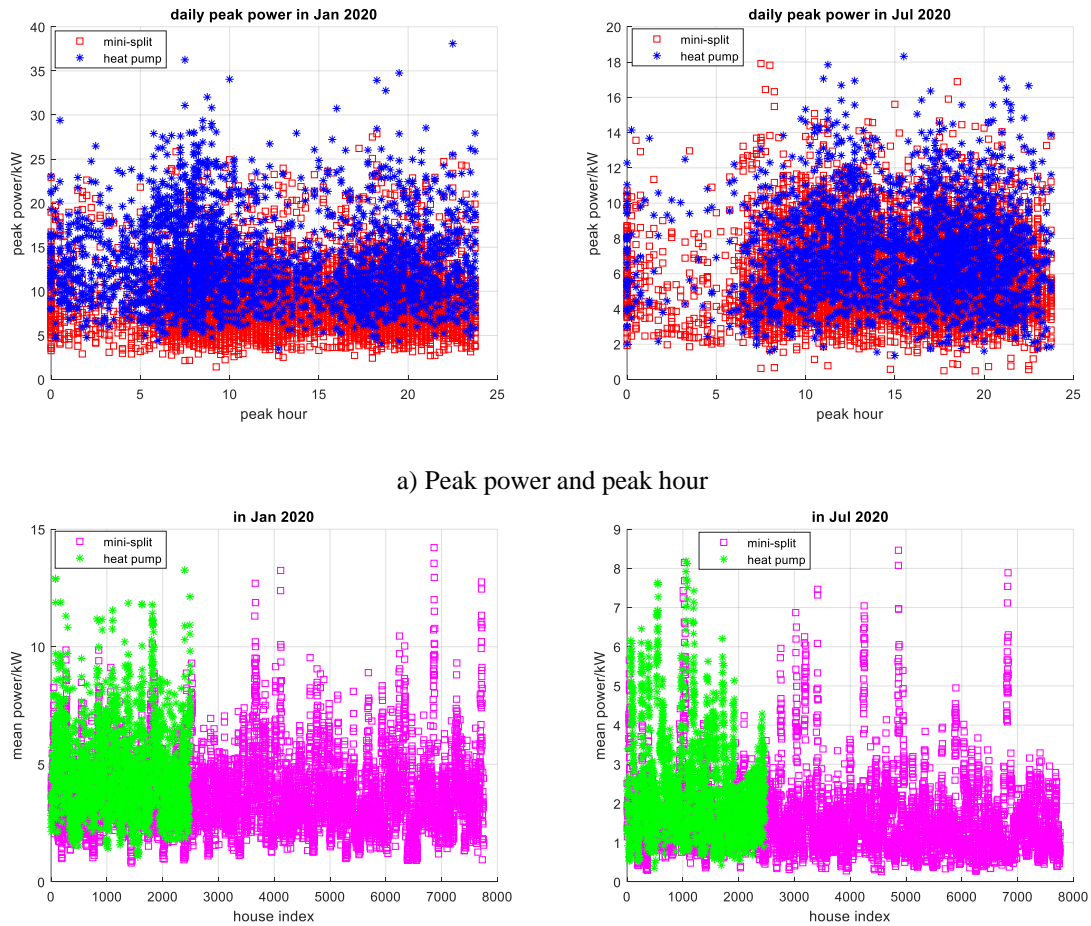
The T-test is used to assess the statistical significance of load profiles differing in shape for the two sources. Fig. 4-8 and Table 4-7 shows that in January, the two sources have a **50.98%** (26/51) probability of load shape difference which indicates it is very hard to distinguish load patterns of these two sources. For July, there is only an **18.33%** (11/60) probability of indicating different load shapes, while an **81.67%** (49/60) probability of similar load shapes. Both are poor results.

Table 4-7 T-test result (MS vs BB)

Month	# of daily shapes	# of available daily shapes	T-test (MS vs BB)			
			Combined greater than/less than probabilities	Equivalent probability	Percent separability	Percent inseparability
Jan-20	16062	15565	26	25	50.98%	49.02%
Jul-20	16108	15598	11	49	18.33%	81.67%

4.3.3 MS vs HP

Finally, a comparison is made between MS vs HP homes. Fig. 4-9 shows relatively clear separability using the metrics of daily peak power for both January and July, but less significance in separability for both months using mean power.



a) Peak power and peak hour

b) Daily mean power for each daily shape

Fig. 4-9 Peak power and mean power (MS vs HP)

All the available load patterns in January and July are grouped into 60 clusters (Fig. 4-2) based on the K-means algorithm, respectively. Plots in Fig 4-10 demonstrate the number of load shapes in each cluster, and show distinguishing characteristics in the cluster index for both January and July between MS and HP.

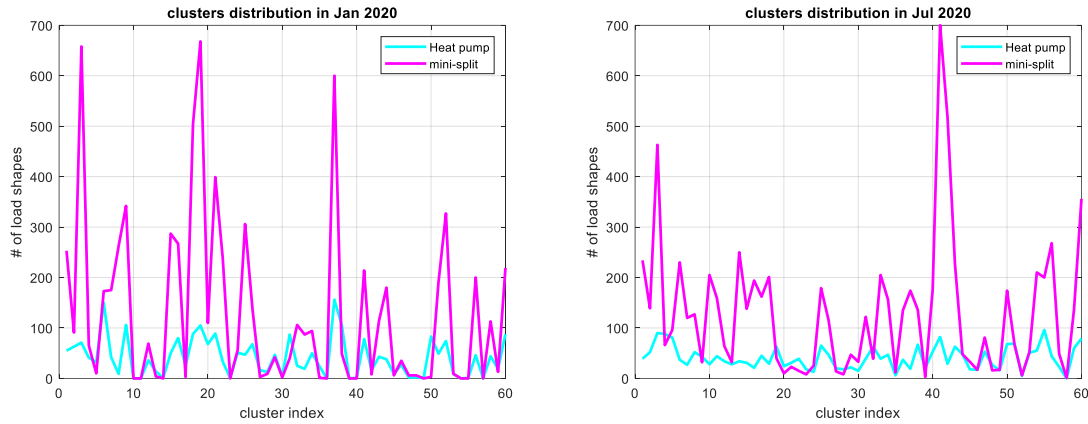


Fig. 4-10 # of load shapes in each cluster

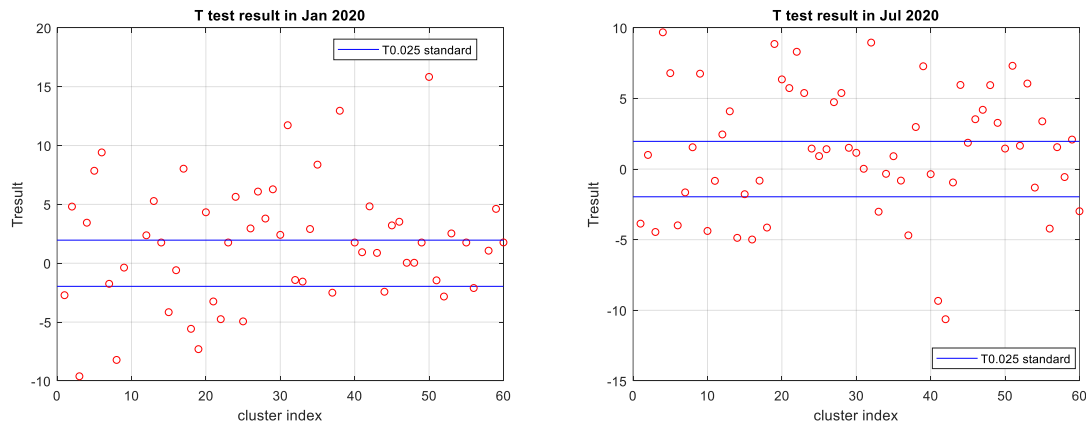


Fig. 4-11 T-test results (MS vs HP)

The T-test is used to assess the statistical significance of load profiles differing in shape for the two sources. Fig. 4-11 and Table 4-8 show that in January the two sources have a **68.52%** (37/54) probability of load shapes being dissimilar which is a fair indicator of different load shapes being represented for these two sources. For July, there is **61.67%**

(37/60) probability of indicating different load shapes. Statistically, these results show homes with HP and MS units to have different load profiles.

Table 4-8 T-test results (MS vs HP)

Month	# of daily shapes	# of available daily shapes	T-test (MS vs HP)			
			Combined greater than/less than probabilities	Equivalent probability	Percent separability	Percent inseparability
Jan-20	16062	15565	37	17	68.52%	31.48%
Jul-20	16108	15598	37	23	61.67%	38.33%

4.4 Electric heating source classification based on ML algorithms

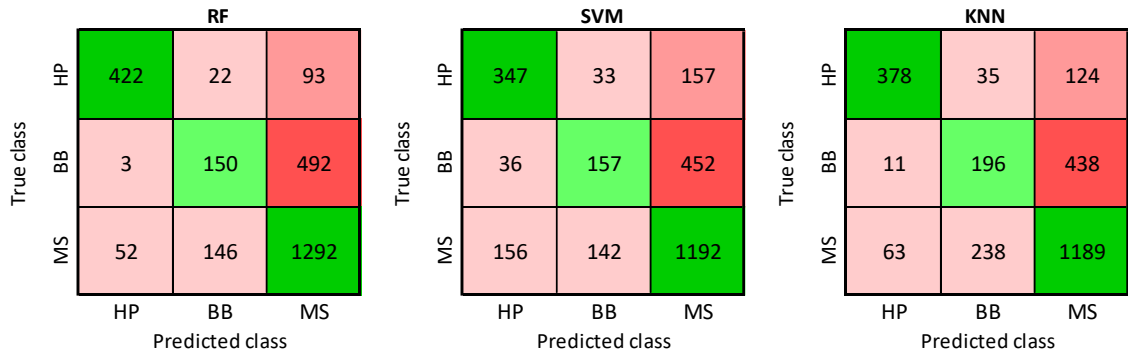
Student’s T-test based analysis is unable to consistently classify the load profiles based electric heating sources during summer and winter months, this objective was extended to use ML-based techniques for training and classification a residence based on the presence of HP, BB or MS. Similar to Chapter 3, RF, SVM and KNN are used to implement the electric heating source classification and have previously been introduced in section 2.5.

There are 448 available smart meters in Dataset 2 for these three electric heating sources, which account for about 90% of the customers. According to the heating analysis in section 4.2.1, winter consumption data in the months of January and February was selected for supervised learning. Three feature groups with different features are chosen for the training and testing of RF, SVM and KNN: Feature Group 1 with 62 consumption and 8 weather features, Feature Group 2 with 62 consumption features only, and Feature Group 3 with top 10 important predictor features found by using the function “predictorImportance” in MATLAB. The full list of features is attached in Appendix C, and the description of each feature is listed in [35]. Grouping of features sets for each ML technique is also given. For these techniques, Dataset 2 was used for the ML training stage, while additional data from 374 homes was received from NB Power for testing and

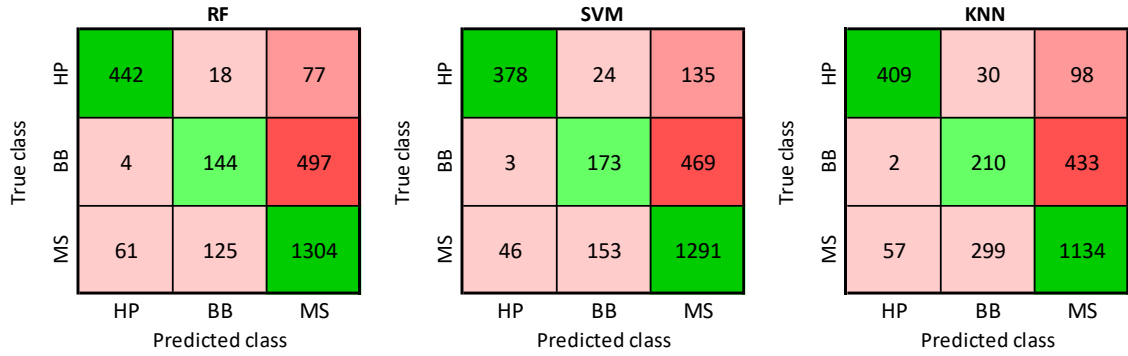
classification results. Both training and testing data are attempted at least 5 times to get an idea of the consistency in classification results.

Fig. 4-12 shows typical classification results using three feature groups with the three ML classification algorithms, respectively. The green color indicates the number of load shapes classified correctly, while the red color shows the number of load shapes that have been categorized to the wrong class. RF results using Feature Group 1 show no obvious difference with results using Feature Group 2 with only uses 62 consumption features. The RF and SVM results using Feature Group 2 generally perform slightly better than Feature Group 1. Generally speaking, adding the weather features does not seem to improve the accuracy of the ML processes for our study. The top 10 important predictors used from Feature Group 3 show consistent results and also do not contain weather features. This is further evidence that weather features are not a key factor in influencing this classification results for these test cases.

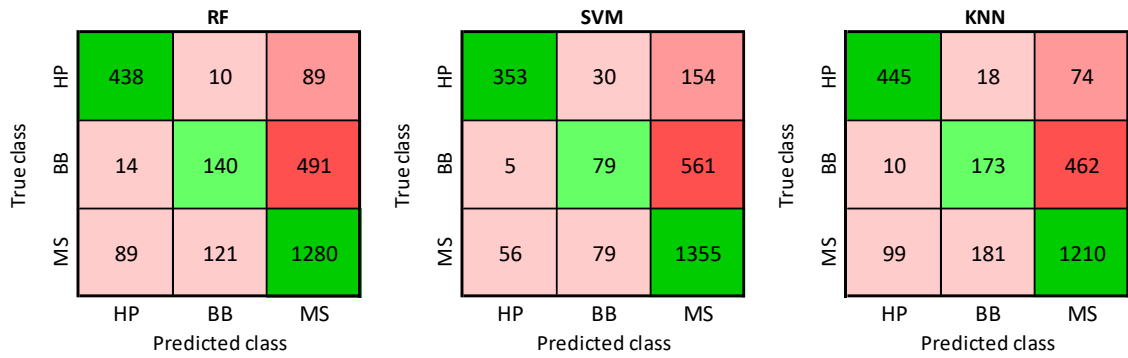
Compared with Feature Group 1 and 2, the accuracy of Feature Group 3 does not degrade much, especially with the RF and KNN algorithms. This is an important result indicating that a full feature set is not necessary to train the ML techniques and that features in Feature Group 3 contribute significantly to the classification results.



a) Results using 70 consumption and weather features



b) Results using 62 consumption features



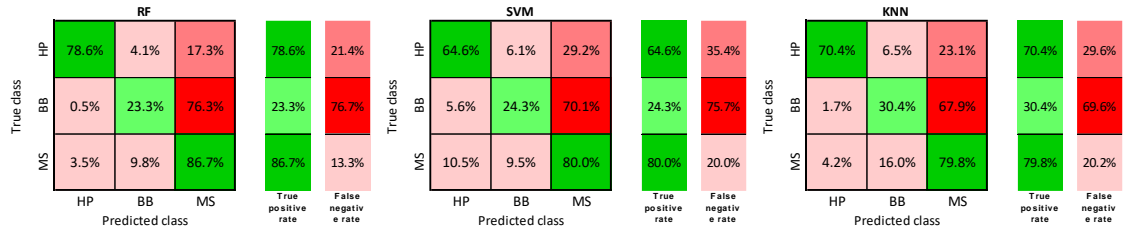
c) Results using the 10 most weighted features

Fig. 4-12 Supervised learning load classification results

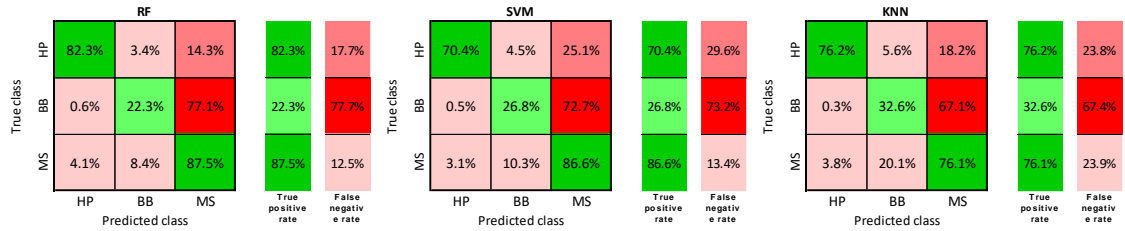
Table 4-9 Consumption data classification results

Groups	Algorithm	HP	BB	MS	Subtotal	Accuracy
#1: Consumption features & weather features (70 features)	RF	422/537	150/645	1292/1490	1864/2672	69.8%
	SVM	347/537	157/645	1192/1490	1696/2672	63.5%
	KNN	378/537	196/645	1189/1490	1763/2672	66.0%
#2: Consumption features (62 features)	RF	442/537	144/645	1304/1490	1890/2672	70.7%
	SVM	378/537	173/645	1291/1490	1842/2672	68.9%
	KNN	409/537	210/645	1134/1490	1753/2672	65.6%
#3: Top 10 important features	RF	438/537	140/645	1280/1490	1858/2672	69.5%
	SVM	353/537	79/645	1355/1490	1787/2672	66.9%
	KNN	445/537	173/645	1210/1490	1828/2672	68.4%

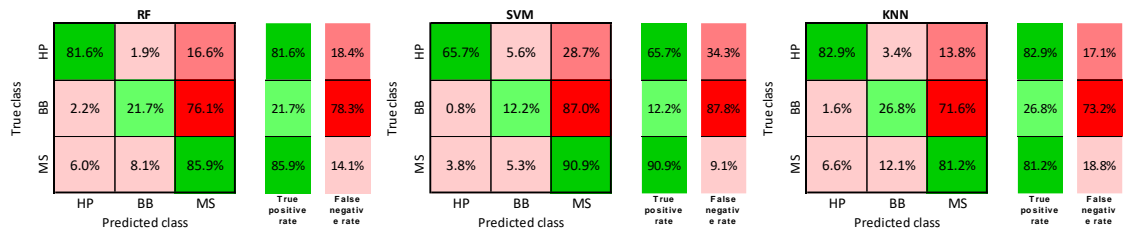
The accuracy of the classification results is shown in Table 4-9. It is seen that the results in Feature Group 2 with the three algorithms outperform Feature Group 1, which means weather features do not seem to improve the classification result. The accuracy of Feature Group 3 shows that the top 10 predictors have the majority of influence on the ML modelling. Overall, RF is generally the top performer when compared to SVM and KNN.



a) Results using 70 consumption and weather features



b) Results using 62 consumption features

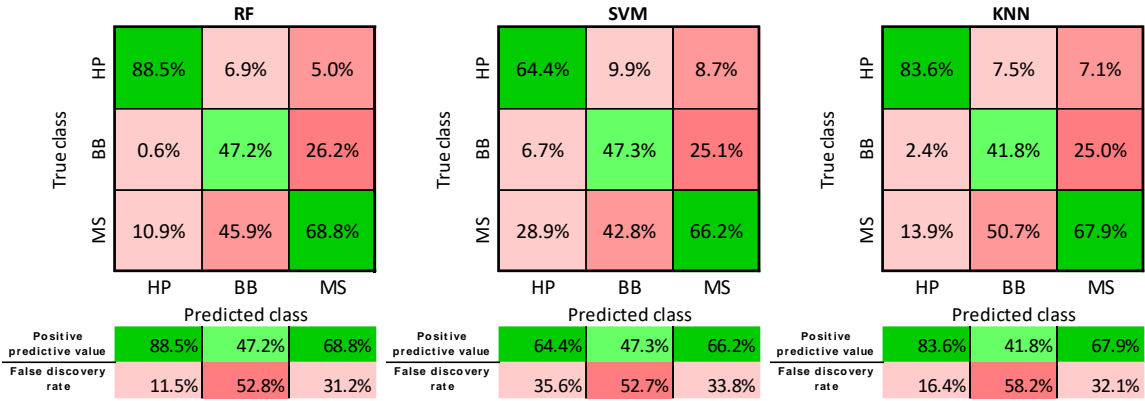


c) Results using the 10 most weighted features

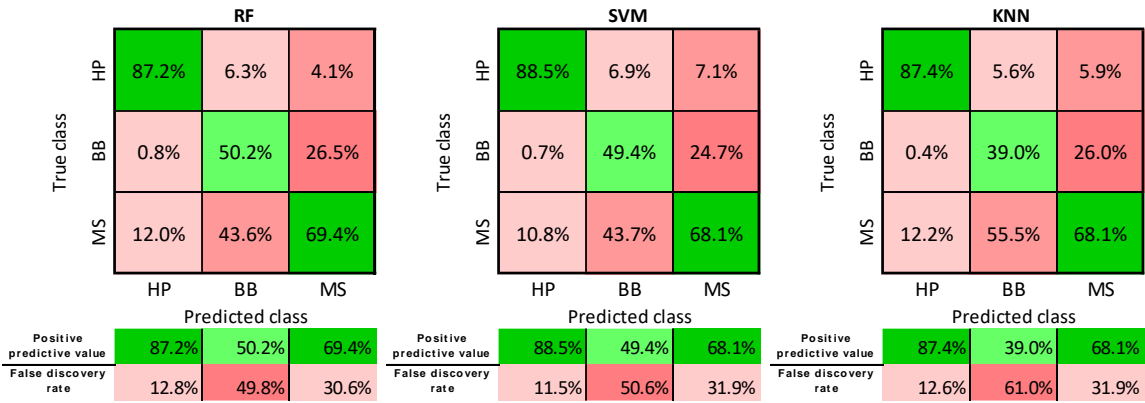
Fig. 4-13 True positive rate and false negative rate of different groups

Fig. 4-13 shows the True Positive Rate (TPR) and False Negative Rate (FNR) of the three groups with the three different supervised learning algorithms, respectively. MS observations have the best TPR, while HP observations have relatively high TPR, and BB observations have a very poor TPR. The results correspond with the analysis in 4.3. The

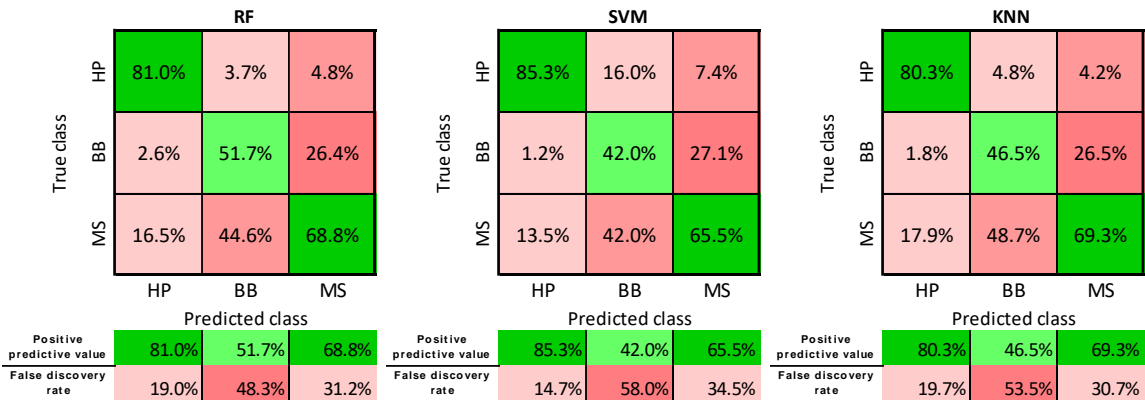
HP has clear discrimination with BB and MS, while the BB and MS are difficult to distinguish, and most of the BB observations are assigned to MS incorrectly.



a) Results using 70 consumption and weather features



b) Results using 62 consumption features



c) Results using the 10 most weighted features

Fig. 4-14 Positive predictive value and false discovery rate of different subset

Fig. 4-14 shows the Positive Predictive Value (PPV), and False Discovery Rate (FDR) of the three ML models, respectively. The predicted classes of HP and MS have a high PPV, while quite a proportion of the miss rates is attributed to the predicted class of BB.

Again it can be seen that adding weather features do not consistently improve the accuracy results and the RF algorithm generally outperforms SVM and KNN algorithms for this classification problem. Comparison of MS and HP have high TPR while BB performs very poorly.

4.5 Analysis and summary

In this chapter, Dataset 2 with the single feature of heating type is introduced to implement load classification and analysis based on the K-means algorithm and Student's T-test. Moreover, supervised ML algorithms of RF, SVM and KNN are used to classify the three most popular electric heating types of HP, BB and MS.

Firstly, load shape classification based on K-means clustering and Student's T-test is applied to the simple test of distinguishing BB vs NE, HP vs NE, and MS vs NE heating sources. The T-test results show that the BB vs NE comparison has significant differences of load shapes in cold days but limited differences in hot days; the HP vs NE comparison has big differences of load shapes both in cold days and hot days; the MS vs NE comparison has a relatively clear boundary of load shapes in cold days but shows limited differences in hot days.

Residential load profiles using the three main electric heating sources (HP, BB, and MS) are assessed using K-means algorithm and Student's T-test. The T-test results show that the HP vs BB comparison has significant differences in load shapes on hot days but no obvious difference in cold days; the MS vs HP comparison has some clear boundaries

of load shapes in both hot days and cold days; while the MS vs BB comparison has no clear differences in load shapes in hot days, and their difference in cold days is also not clear.

Since the Student T-test analysis was shown to have limited effectiveness in separating electric heating sources for cold winter days and hot summer days, RF, SVM and KNN were then used to classify load profiles based on electric heating sources. The result shows that RF has the highest accuracy of classification result, although the differences were slight. Among these three heating types HP and MS have a relatively high True Positive Rate, indicating separability between these two heating sources, while the ability to detect BB is quite poor. Comparison among the three feature groups, Feature Group 3 using the top 10 weighted features gets relatively high accuracy without using a large feature set. Overall, Feature Group 2 with only consumption features has the highest accuracy, which means for our test case, the weather features have limited contributions to achieving higher classification accuracies.

CHAPTER 5

5 CONCLUSION

The main research goal of this project is to discriminate residential load profiles based on limited consumption data and associated metadata. Two sets of consumption data were provided by NB Power; the first data set contains three metadata features: property value, lot size and house type, while the second dataset from Shediac NB has only a single metadata feature of heating source.

In this research, three general methodologies were utilized to accomplish the classification of energy consumption data from Dataset 1. The first approach used simple metrics such as PMR, average daily consumption and peak power are used to separate residential profiles based on particular metadata features. It was found that such individual features had limited ability to perform this task given the significant impact of unknown heating types of each home. By combining certain metadata features such as lot size and home price and utilizing more sophisticated statistical techniques forms the second approach. This approach had limited success in summer months but had a more difficulties separating load profiles in winter months. The third approach for this assessment used ML algorithms and better performance was found compared to the second approach. For Dataset 2, the focus was on load classification based on heating type and again three assessment approaches were utilized. Simple metrics, statistical methods, along with three ML algorithms (RF, SVM, and KNN), were used to differentiate homes based detected electric heating source type. As with Dataset 1, both simple metrics and statistical methods were not able to consistently determine heating source type (BB, MS or HP) based on the consumption data. ML techniques resulted in better results, but the BB case still proved

problematic. Studies were then performed to assess the ML performance with a reduced feature group comprised of the top 10 weighted features assessed during the training and validation stage of the ML process. It was found that this reduced feature group performed almost on par with the large feature group, which is useful in reducing the computational load of the training and classification phase.

5.1 Further comments of Dataset 1

Two groups of energy daily consumption were built based on combining metadata features; large lot size with more expensive homes (Group 1), while small lot sizes with low prices forms (Group 2). It was hoped to create a situation of greater separation of cluster centroids with the assumption that Group 1 and 2 represent larger and smaller sized homes, respectively. Clustering and the student's T-test shows that these two groups can have significant differences of load shapes on hot summer days while much less significant differences in cold winter days. Cold days present a more difficult problem for classification given the unknown heating type and home efficiency. Cooling in summer seems to provide enough discernable information to indicate a statistical difference. Perhaps this is due to the differences in the use of air-conditioning for Groups 1 and 2, although cannot be known with certainty.

Three ML techniques (RF, SVM and KNN) were then applied to Dataset 1 to see if improvement could be made on results found using statistical analysis. These results show that both winter months and summer months can be classified based on Group 1 and 2 with very high accuracy, while the accuracy in summer months is higher than in winter months. Consequently, ML techniques are able to achieve separability where the statistical methods

failed. Data limitations, however, should be taken into consideration since more diverse consumption data for training and testing will inevitably lead to poorer accuracy results.

5.2 Further comments of Dataset 2

Characteristics of different heating types are analyzed for Dataset 2. Three of the most common electric heating sources (HP, BB, and MS) are chosen to differentiate residential load profiles. Once again K-means and student's T-test are used with limited success as it was found that BB heating is difficult distinguish between MS and HPs. This is important given the dominance of BB heating in New Brunswick.

ML algorithms are then introduced to implement the load classification of these three heating types. Supervised learning algorithms of RF, SVM and KNN are used to classify the load shapes based on the detection of HP, BB, and MS in the residential unit. The results show that HP and MS have a high true positive rate, while isolating BB heating is difficult. That said, the ML methods were found to be more robust and consistent in determining differences between MS and HP units in comparison to the statistical method.

5.3 Future work

Future work and research should include:

- Larger and more diverse consumption data is needed to ensure that conclusions drawn from the analysis of Dataset 1 and 2 are valid.
- Utilizing other metadata features such as occupancy, appliances, behavior, family income, etc., which can also affect residential energy consumption, but a lack of data has prevented such an application.
- Investigating other ML techniques to detect the presence of BB heating units when homes are known to use an electric heating source. This will help utilities

understand the extent of more efficient heating and cooling technologies being installed or in use.

- Determining, if available, improvements in residential classification and heat source detection when using higher resolution temporal consumption data, perhaps moving to 1-minute time resolution rather the 15-minute data used in this thesis.
- A better understanding of the impact of weather on residential load profiles. Work in this thesis showed weather to have limited impact of classification accuracy but common-sense dictates that this should not be the case.

Bibliography

- [1] NB Power, “Load Forecast 2020-2030”, 2019
- [2] Adrian Albert and Ram Rajagopal, “Smart meter driven segmentation- what your consumption says about you”, IEEE Trans. Power Syst., Vol. 28, No. 4, Nov 2013
- [3] Jungsuk Kwac, June Flora, and Ram Rajagopal, “Household energy consumption segmentation using hourly data”, IEEE Trans. on Smart Grid, Vol. 5, No.1, Jan 2014
- [4] Guardian News and Media Limited, “A guardian guide to your metadata”. theguardian.com. 12 June 2013. Archived from the original on 6 March 2016.
- [5] C. S. Chen, M. S. Kang, T.Y. Yo, and C.W. Huang, “Synthesis of system power profile and temperature sensitivity analysis”, 2003 IEEE Bologna PowerTech Conference, June 23-26, Bologna, Italy
- [6] Jingchao Zhang, Anhe Yan, Zhuoya Chen and Kun Gao, “Dynamic synthesis load modeling approach based on load survey and load curves analysis”, DRPT2008, 6-9 April 2008, Nanjing, China
- [7] A. Rosin, H. Hõimoja, T. Möller, M. Lehtla, “Residential electricity consumption and loads pattern analysis”, 2010 Electric Power Quality and Supply Reliability Conference
- [8] Poppy Rowena Harvey, Bruce Stephen, Stuart Galloway, “Classification of AMI residential load profiles in the presence of missing data”, IEEE Trans. on Smart Grid, Vol.7, No.4, July 2016
- [9] Xiaouou Monica Zhang, Katarina Grolinger, Miriam A. M. Capretz, Luke Seewald, “Forecasting residential energy consumption: single household perspective”, 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)
- [10] Claes Sandels, Marcus Kempe, Magnus Brodin, Anders Mannikoff, “Clustering residential customers with smart meter data using a data analytic approach – External validation and robustness analysis”, 2019 9th International Conference on Power and Energy Systems (ICPES)
- [11] Shama Naz Islam, Akhlaqur Rahman, Lawrence Robinson “Load profile segmentation using residential energy consumption data”, 2020 International Conference on Smart Grids and Energy Systems (SGES)
- [12] Hyeob Yu, Jin Ki Lee, Jong Min Ko and Sun Ic Kim. “A method for classification of electricity demands using load profile data”, the Fourth Annual ACIS International Conference on Computer and Information Science (ICIS’05)
- [13] G. Chicco et al., “Load pattern-based classification of electricity customers,” IEEE Trans. Power Syst., Vol. 19, No. 2, pp. 1232–1239, May 2004

- [14] G. Chicco, R. Napoli, F. Piglione, “Comparisons among clustering techniques for electricity customer classification”, IEEE Trans. Power Syst., Vol. 21, No. 2, May 2006
- [15] S. Bhatia, “Adaptive K-means clustering,” in Proc. Int. Florida Artif. Intell. Res. Soc. Conf., 2004
- [16] Guo-ying Lin, Feng Pan, Yu-yao Yang, Lin Yang, Guangyu He, Shuai Fan “The pattern recognition of residential power consumption based on HMM”, 2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)
- [17] Rui Lang, Kaiyan Wang, Jiao Wang, Rong Jia, “Analysis of residents' differentiated power consumption behavior based on load classification”, 2019 12th International Symposium on Computational Intelligence and Design (ISCID)
- [18] Gheorghe Grigoras, Ovidiu Ivanov, Mihai Gavrilas, “Customer classification and load profiling using data from smart meters”, 2014 12th Symposium on Neural Network Applications in Electrical Engineering (NEUREL)
- [19] Mohamed Chaouch, “Clustering-based improvement of nonparametric functional time series forecasting: application to intra-day household-level load curves”, IEEE Trans. on Smart Grid, Vol. 5, No. 1, Jan 2014
- [20] Mingyang Sun, Ioannis Konstantelos, Goran Strbac, “C-vine copula mixture model for clustering of residential electrical load pattern data”, IEEE Trans. Power Syst., Vol. 32, No. 3, May 2017
- [21] Mengqiu Fang, Yue Xiang, Li Pan, Bohan Xu, Youbo Liu, Junyong Liu, Tianhao Wang, “Data-driven load pattern identification”, 2021 IEEE IAS Industrial and Commercial Power System Asia
- [22] Inam Ullah Khan, Nadeem Javaid, C. James Taylor, Kelum A. A. Gamage, Xiandong Ma, “Big data analytics based short term load forecasting model for residential buildings in smart grids”, IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)
- [23] J Jeyaranjani, D Devaraj, “Deep learning based smart meter data analytics for electricity load prediction”, 2019 IEEE International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development (INCCES)
- [24] Heng Shi, Minghao Xu, Ran Li, “Deep learning for household load forecasting--a novel pooling deep RNN”, IEEE Trans. on Smart Grid, Vol. 9, No. 5, Sept 2018
- [25] Javaria Hameed, Rabiya Khalid, Muhammad Umar Javed, “Enhanced classification with logistic regression for short term price and load forecasting in smart homes” 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)

- [26] Shiyin Zhong, Kwa-Sur Tam, “Hierarchical classification of load profiles based on their characteristic attributes in frequency domain”, *IEEE Trans. Power Syst.*, Vol. 30, No. 5, Sept 2015
- [27] Baran Yildiz, Jose I. Bilbao, Jonathon Dore, Alistair Sproul, “Household electricity load forecasting using historical smart meter data with clustering and classification techniques”, 2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)
- [28] Kushan A. Choksi, Sachin K. Jain, “Novel computational-index as a representative feature for non-intrusive load monitoring”, *Proceedings of 2018 IEEE International Conference on Information Communication and Signal Processing (ICSP 2018)*
- [29] Ramon Granell, Colin J. Axon, David C. H. Wallom, “Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles”, *IEEE Trans. Power Syst.*, Vol. 30, No. 6, Nov 2015
- [30] Xinan Wang, Yishen Wang, Jianhui Wang, Di Shi, “Residential customer baseline load estimation using stacked autoencoder with pseudo-load selection”, *IEEE Journal on Selected Areas in Communications*, Vol. 38, No. 1, Jan 2020
- [31] Po Hu, Yunjia Wang, Ning Pang, Zeya Zhang, Yifan Huang, Haoran Guo, “Research on residential load classification method based on multi-model parallel integration algorithm”, 2021 International Conference on Cyber-Physical Social Intelligence (ICCSI)
- [32] Mahmood Reaz Sunny, Md Ahsan Kabir, Roubaiath Islam, Saraban Nazifa, “Smart meter data compression and load profile classification using UMAP and random forest”, 2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)
- [33] Ailin Asadinejad, Kevin Tomsovic, Chien-fei Chen, “Impact of residential customer classification on demand response results under high renewable penetration”, 2017 IEEE Power & Energy Society General Meeting
- [34] V. Singhal, J. Maggu, and A. Majumdar, “Simultaneous detection of multiple appliances from smart-meter measurements via multi-label consistent deep dictionary learning and deep transform learning,” *IEEE Trans. on Smart Grid*, Vol. 10, No. 3, pp. 2969–2978, 2019.
- [35] Konstantin Hopf, Mariya Sodenkamp, Thorsten Staake. “Enhancing energy efficiency in the residential sector with smart meter data analytics”. *Electronic Markets* (2018) 28:453 – 473
- [36] Package “SmartMeterAnalytics”. August 18, 2020
- [37] Andreas Weigert, Konstantin Hopf, Nicolai Weinig and Thorsten Staake. “Detection of heat pumps from smart meter and open data”. *Energy Informatics* 2020, 3(Suppl 1):21

- [38] H Fei, Y Kim, Sambit Sahu, Milind Naphade. "Heat pump detection from coarse grained smart meter data with positive and unlabeled learning", KDD'13, August 11–14, 2013, Chicago, Illinois, USA
- [39] Forgy, Edward W. (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*. 21 (3): 768–769. JSTOR 2528559
- [40] Kriegel, Hans-Peter, Schubert, Erich, Zimek, Arthur (2016). "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?". *Knowledge and Information Systems*. 52 (2): 341–378. ISSN 0219-1377.
- [41] MacKay, David (2003). "Chapter 20. An example inference task: Clustering". *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. pp. 284–292. ISBN 978-0-521-64298-9. MR 2012999.
- [42] Hamerly, Greg, Elkan, Charles (2002). "Alternatives to the k-means algorithm that find better clusterings". *Proceedings of the eleventh international conference on Information and knowledge management (CIKM)*.
- [43] San Jose State University. "6: Introduction to null hypothesis significance testing."
- [44] Goulden, C. H. "Methods of statistical analysis", 2nd ed. New York: Wiley, pp. 50-55, 1956.
- [45] "Student" William Sealy Gosset (1908). "The probable error of a mean". *Biometrika*. 6 (1): 1–25.
- [46] Pfanzagl J, Sheynin O (1996). "Studies in the history of probability and statistics. XLIV. A forerunner of the t-distribution". *Biometrika*. 83 (4): 891–898.
- [47] "T Table", <https://www.tdistributiontable.com/>
- [48] Tony Yiu, "Understanding random forest". *Towards Data Science*, Jun 2019
- [49] Sruthi E R, "Understanding random forest", *Analytics Vidhya*, Jun 2021
- [50] Rohith Gandhi, "Support vector machine introduction to machine learning algorithms". *Towards Data Science*, Jun 2018
- [51] Cortes, Corinna, Vapnik, Vladimir (1995). "Support-vector networks". *Machine Learning*. 20 (3): 273–297.
- [52] Sunil Ray, "Understanding support vector machine (SVM) algorithm from examples (along with code)". *Analytics Vidhya*, Sept 13, 2017
- [53] M Hussain, S Wajid, A Elzaart, M Berbar, "A comparison of SVM kernel functions for breast cancer detection". *2011 8th International Conference Computer Graphics, Imaging and Visualization*

- [54] Baeldung, “Multiclass classification using support vector machines”, Baeldung CS, Nov 11, 2022
- [55] Onel Harrison, “Machine learning basics with the k-nearest neighbors algorithm”, Towards Data Science, Sept 2018
- [56] Tavish Srivastava, “Introduction to k-nearest neighbors: A powerful machine learning algorithm (with implementation in Python & R)”, Analytics Vidhya, Mar 2018
- [57] <https://www.mathworks.com/help/stats/compactregressionensemble.predictorimportance.html>
- [58] Powers, David M. W. (2011). “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation”. *Journal of Machine Learning Technologies*. 2 (1): 37–63.
- [59] Tharwat A. (August 2018). “Classification assessment methods”. *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003.
- [60] “How efficient are heat pumps compared to baseboard heaters?” <https://takechargenl.ca/faq/how-efficient-are-heat-pumps-compared-to-baseboard-heaters/>

Appendix A

Table A T-distribution values [47]

t Table											
cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Appendix B

Table-B Full list of features used in Dataset 1 with the accuracy of different supervised learning algorithms

Index	Features	winter	summer
1	c15_week	X	X
2	s15_min	X	X
3	s15_max	X	X
4	r15_mean_max	X	X
5	r15_min_mean	X	X
6	s15_we_max	X	X
7	s15_we_min	X	X
8	s15_wd_max	X	X
9	s15_wd_min	X	X
10	r15_min_wd_we	X	X
11	r15_max_wd_we	X	X
12	s15_q1	X	X
13	s15_q2	X	X
14	s15_q3	X	X
15	s15_min_avg	X	X
16	s15_max_avg	X	X
17	s15_variance	X	X
18	s15_var_we	X	X
19	s15_var_wd	X	X
20	r15_var_wd_we	X	X
21	s15_cor	X	X
22	s15_cor_we	X	X
23	s15_cor_wd	X	X
24	s15_cor_wd_we	X	X
25	s15_sm_variety	X	X
26	s15_bg_variety	X	X
27	s15_sm_max	X	X
28	s15_number_zeros	X	X
29	c15_evening_no_min	X	X
30	c15_morning_no_min	X	X
31	c15_night_no_min	X	X
32	c15_noon_no_min	X	X
33	c15_afternoon_no_min	X	X

34	r15_mean_max_no_min	X	X
35	r15_evening_noon_no_min	X	X
36	r15_morning_noon_no_min	X	X
37	r15_day_night_no_min	X	X
38	t15_above_0.5kw	X	X
39	t15_above_1kw	X	X
40	t15_above_2kw	X	X
41	t15_above_mean	X	X
42	t15_daily_max	X	X
43	t15_daily_min	X	X
44	s15_num_peaks	X	X
45	s15_diff	X	X
46	ts15_acf_mean3h	X	X
47	ts15_acf_mean3h_weekday	X	X
48	ts15_stl_varRem	X	X
49	t15_above_base	X	X
50	s15_day_diff	X	X
51	s15_day_diff_weak	X	X
52	t15_wide_peaks	X	X
53	t15_width_peaks	X	X
54	t15_time_above_base2	X	X
55	t15_percent_above_base	X	X
56	t15_value_above_base	X	X
57	t15_const_time	X	X
58	t15_value_min_guess	X	X
59	t15_first_above_base	X	X
60	s15_num_big_peaks	X	X
61	t15_number_small_peaks	X	X
62	t15_dist_big_v	X	X
RF	Accuracy (Training)	90.9%	97.9%
RF	Accuracy (Test)	83.3%	100.0%

SVM	Accuracy (Training)	83.1%	93.7%
SVM	Accuracy (Test)	83.3%	97.7%

KNN	Accuracy (Training)	75.3%	89.5%
KNN	Accuracy (Test)	85.7%	93.0%

Appendix C

Table-C Full list of features used in Dataset 2 with the accuracy of different supervised learning algorithms

Index	Features	consumption & weather features	consumption features	Top 10 weighted features
1	cor_overall	X		
2	cor_daily	X		
3	cor_night	X		
4	cor_daytime	X		
5	cor_evening	X		
6	cor_minima	X		
7	cor_maxmin	X		
8	cor_weekday_weekend	X		
9	c15_week	X	X	
10	s15_min	X	X	
11	s15_max	X	X	X
12	r15_mean_max	X	X	
13	r15_min_mean	X	X	X
14	s15_we_max	X	X	
15	s15_we_min	X	X	
16	s15_wd_max	X	X	
17	s15_wd_min	X	X	
18	r15_min_wd_we	X	X	
19	r15_max_wd_we	X	X	
20	s15_q1	X	X	
21	s15_q2	X	X	
22	s15_q3	X	X	
23	s15_min_avg	X	X	
24	s15_max_avg	X	X	
25	s15_variance	X	X	X
26	s15_var_we	X	X	X
27	s15_var_wd	X	X	X
28	r15_var_wd_we	X	X	
29	s15_cor	X	X	
30	s15_cor_we	X	X	
31	s15_cor_wd	X	X	
32	s15_cor_wd_we	X	X	

33	s15_sm_variety	X	X	
34	s15_bg_variety	X	X	X
35	s15_sm_max	X	X	
36	s15_number_zeros	X	X	
37	c15_evening_no_min	X	X	
38	c15_morning_no_min	X	X	
39	c15_night_no_min	X	X	X
40	c15_noon_no_min	X	X	
41	c15_afternoon_no_min	X	X	
42	r15_mean_max_no_min	X	X	X
43	r15_evening_noon_no_min	X	X	
44	r15_morning_noon_no_min	X	X	
45	r15_day_night_no_min	X	X	
46	t15_above_0.5kw	X	X	
47	t15_above_1kw	X	X	
48	t15_above_2kw	X	X	
49	t15_above_mean	X	X	
50	t15_daily_max	X	X	
51	t15_daily_min	X	X	
52	s15_num_peaks	X	X	
53	s15_diff	X	X	
54	ts15_acf_mean3h	X	X	
55	ts15_acf_mean3h_weekday	X	X	
56	ts15_stl_varRem	X	X	X
57	t15_above_base	X	X	X
58	s15_day_diff	X	X	
59	s15_day_diff_weak	X	X	
60	t15_wide_peaks	X	X	
61	t15_width_peaks	X	X	
62	t15_time_above_base2	X	X	
63	t15_percent_above_base	X	X	
64	t15_value_above_base	X	X	
65	t15_const_time	X	X	
66	t15_value_min_guess	X	X	
67	t15_first_above_base	X	X	
68	s15_num_big_peaks	X	X	
69	t15_number_small_peaks	X	X	
70	t15_dist_big_v	X	X	
RF	Accuracy (Training)	73.6%	73.5%	72.1%
RF	Accuracy (Test)	69.8%	70.7%	69.5%

SVM	Accuracy (Training)	71.5%	72.0%	69.5%
SVM	Accuracy (Test)	63.5%	68.9%	66.9%

KNN	Accuracy (Training)	70.0%	71.1%	70.6%
KNN	Accuracy (Test)	66.0%	65.6%	68.4%

Curriculum Vitae

Candidate's full name: Chunjin Liu

Universities attended (with dates and degrees obtained):

Beijing Jiaotong University, China, Bachelor of Electrical Engineering, 2007

Beijing Jiaotong University, China, Master of Power Electronics and Power Drive, 2009

Publications: N/A

Conference Presentations: N/A