5-2017

A data-driven approach for benchmarking energy efficiency of warehouse buildings

Wee Leong LEE
Singapore Management University, wllee@smu.edu.sg

Kar Way TAN
Singapore Management University, kwtan@smu.edu.sg

Zui Young LIM
Singapore Management University, zylim.2011@sis.smu.edu.sg

A Data-Driven Approach for Benchmarking Energy Efficiency of Warehouse Buildings

Wee Leong Lee, Kar Way Tan, and Zui Young Lim

Abstract — **This study proposes a data-driven approach for benchmarking energy efficiency of warehouse buildings. Our proposed approach provides an alternative to the limitation of existing benchmarking approaches where a theoretical energy-efficient warehouse was used as a reference. Our approach starts by defining the questions needed to capture the characteristics of warehouses relating to energy consumption. Using an existing data set of warehouse building containing various attributes, we first cluster them into groups by their characteristics. The warehouses characteristics derived from the cluster assignments along with their past annual energy consumption are subsequently used to train a decision tree model. The decision tree provides a classification of what factors contribute to different levels of energy consumption. Finally, we showed how a linear regression method is used to predict the energy consumption based on relationships between strongly correlated variables, such as climate zone, number of working hours, and floor area. With our proposed data-driven approach, decision makers can analyze and benchmark their warehouse building data, adopt best practices from existing solutions and make better decisions when recommending high-impact energy reduction solutions for their warehouses.**

I. Introduction

The study of energy efficiency in buildings has become an important topic in environmental conservation as governments and corporations alike recognize the priority of curbing greenhouse emissions and improving environmental sustainability in their operations [1]. In the context of manufacturing and supply chain industries, appropriate energy management in warehouses has significant impact on the operating cost. To evaluate energy efficiencies of warehouse, benchmarking is often performed. The task of benchmarking energy efficiency is a complex process of collecting, analyzing, and scoring data. A key challenge faced by data analysts and business managers is knowing which factors are most crucial when measuring energy efficiency? In numerous studies we surveyed, a combination of inputs from experts, past building characteristics and energy data are used to develop a theoretical model of an "ideal" energy-efficient warehouse, to be used as a benchmarking model. The reports from such benchmarking exercise usually provide qualitative analysis of the energy efficiency performance of the studied warehouse building against all other buildings used to build the "ideal" benchmark warehouse. As each warehouse has its unique climate, operational and building characteristics, which limits implementable measures suggested from such benchmarking exercises, it is difficult for decision makers to gain actionable insights from a generic benchmark results.

Our proposed data-driven approach provides a way to characterize warehouse buildings to benchmark and predict warehouses expected energy consumption based on the results of other buildings with similar characteristics. We demonstrated that similarities between buildings can be determined using a clustering technique. Information about each warehouse can then be summarized and grouped into clusters. The cluster groups are found to be strongly correlated to energy consumption. A decision tree model is subsequently used to classify the cluster groups to predict the likely range of energy consumption of each warehouse based on the set of similar warehouses. Decision trees provide interpretable results and help decision makers identify common characteristics between groups of warehouses, which operate within an energy consumption range, thus providing opportunities for improvements based on the characteristics relevant to the studied warehouse. Although our study was based on a specific set of warehouse buildings, we believe that the methodology can be applied to another set of building data to yield similar results. The application of the proposed approach is particularly relevant to organizations managing large number of buildings at different locations, especially in the manufacturing and logistics industry. Data can be collected within the company across different geographical locations or combined with publicly available building data.

Our contributions are in two-folds. Firstly, we provided a practical approach, using data mining methods to benchmarking energy efficiencies of buildings by comparing warehouses of similar characteristics. This approach allows decision-makers to better identify characteristics and best practices for improving energy consumptions. Secondly, we provided an additional linear metric that can potentially improve the prediction accuracy of the energy consumption band of warehouses.

II. Literature Review

There exist numerous methods for benchmarking energy efficiency of buildings [2, 3, 4]. Such methods compare the warehouses around a defined "ideal" energy-efficient warehouse. The "ideal" warehouse is defined based on past studies and by experts who build analytical models that define sustainable warehouse operations. These defined rules sets a baseline of what an energy-efficient warehouse is, and the exhibited characteristics of low-emissions or sustainable

*This work was supported by the Green Transformation Lab (a collaboration between DHL and Singapore Management University).

Wee Leong Lee & Kar Way Tan are faculty members with the School of Information Systems, Singapore Management University. Zui Young Lim was a post-graduate master student with the School of Information Systems, Singapore Management University.

warehouses [5]. Guides to sustainable operations such as the Energy Consumption Guide 19 (ECON 19), part of the UK Government's Energy Efficiency Best Practice programme, and US Environmental Protection Agency's ENERGY STAR Best Practices Checklist provide a checklist method to evaluating the environmental sustainability of building operations.

Other available research propose benchmarking buildings based on statistical model, which defines a theoretical efficient frontier. These methods include deriving warehouse performance against the efficient frontier, with methods such as Data Envelopment Analysis (DEA), Stochastic Frontier Analysis (SFA) models [6, 7, 8], or scoring performance using multi-linear regression models such as the ENERGY STAR Portfolio Manager and ENERGY STAR Score for Warehouses in the United States [9, 10].

The above mentioned methods derive benchmarking results by defining a fixed set of parameters and modelling the performance of an ideal warehouse. The benchmark scores for warehouses are computed based on a mathematical model of hundreds of parameters, which describes warehouses operating in diverse operating conditions. The challenge of such an approach is that the results are compared with large range of warehouses, hence difficult to find identifiable characteristics of a 'role model' to gain actionable insights.

Based on the existing literature, benchmarking of warehouses' operating efficiencies are known to depend on a variety of factors, including:

- Energy consumption requirements based on external climatic conditions
- Technological upgrades that have been incorporated into the building's physical design
- Operating activity (from the employees and energy-consuming equipment) during certain periods and throughout the year
- Margin of error when collecting and benchmarking the data collected

In the following sections, we will present our data-driven data mining approach to cluster warehouses based on their common characteristics, identify strongly correlated characteristics to energy consumption and build a predictive model to help decision-makers find opportunities for energy consumption improvements.

III. Data set and Features

The data set used in our study was taken from the U.S. Energy Information Administration (EIA) 2003 Commercial Buildings Energy Consumption Survey (CBECS). The most recent survey data published was from 2003 and 2012. Unfortunately, the data from 2012 survey was incomplete and was omitted from the study. We believe the methodology proposed in this paper can be applied to more recent data set when it becomes available. The survey data can be obtained at http://www.eia.gov/consumption/commercial/data/2003/.

There are 473 unrefrigerated warehouses within the data set of approximately 5,000 buildings. As the energy requirements for refrigerated warehouses are very different, we will only focus on unrefrigerated warehouses in our study. Among the selected warehouses, there are 50 warehouses with missing energy consumption values, which will be omitted in this study, leaving the remaining 423 warehouses for the analysis. From the data set, 81 binary features were separated into 10 categories termed as 'Cluster Groups' according to its description given below:

1. Fuel Sources
2. Physical Building Characteristics
3. Building Improvements
4. Warehouse Ownership
5. Utility of Building
6. Use of Energy
7. Multi-complex Building
8. Warehouse Production of Various Utilities
9. Detailed Heating and Cooling Sources
10. Energy Management Systems

Other variables utilized in our analysis include:

- Energy Consumption in BTU-Equivalent units, which is an aggregate of the Annual Major Fuel Consumption of a warehouse.
- Annual Electricity Consumption, Natural Gas Consumption, District Heat Consumption, Other Fuel Sources Consumption.
- Environment Climate variables (Heating and Cooling degree days to base 65 degree Fahrenheit, Climate Region)
- Operational Characteristics (Number of Working Hours per Week, Number of Employees)
- Equipment (Number of Personal Computers, Servers)
- Heating and Cooling Sources (Percentage Heating and Cooling for each Heating and Cooling Equipment)
- Floor area of the building (in square feet)

IV. Approach

Figure 1 summarizes the iterative approach to carrying out the data analysis. In stage one, we re-evaluated the definition of what constitutes energy efficiency. The benchmarking of energy efficiency can be achieved by applying a framework which analyzes data about physical building characteristics, building energy efficiency improvements, energy consuming fixtures and equipment (such as lighting, servers, computers, etc.), external climate conditions (which affect heating and cooling requirements), building activity levels (such as the number of working hours and number of employees), and energy consumption sources.

In stage two, the questionnaire to collect the response is designed. In this stage, the types of responses to be collected is considered. Continuous variable responses are collected for questions expressed as fractions or proportions, and variables where the upper-bound is not known. Questions such as "percentage of heating by package heaters" and "floor area of building" are examples that should be collected as continuous variables. Continuous variables can then be "binned" or discretized to find boundaries which separate different classes of warehouse with similarities. Discrete variables should be collected for categorical and binary responses. Binary responses are useful when detailed responses are not required or in cases where it is difficult to obtain an accurate response.

Multiple categorical variables can then be used to find clusters of warehouses with similar characteristics.

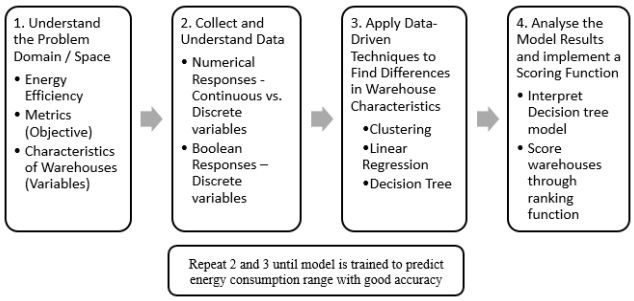

Figure 1. Iterative Approach to executing data analysis on warehouse building data

In stage three, a combination of data mining and data analysis tools are used to analyze the responses and build a predictive model that benchmarks energy consumption among different sets of warehouses with similar characteristics. In this stage, both cluster analysis model and decision tree model are used to analyze characteristics of warehouses for benchmarking. Linear regression analysis is also used to derive new features to better predict energy consumption. Stage four consists of interpretation of results and scoring of the warehouses against others with similar characteristics to draw more meaningful comparisons and recommendations. Stage three and four are not commonly practiced in existing benchmarking methods which are essential to aid decision-makers identify the significant features that impact energy consumptions. The three data-mining techniques used in stage three are explained as follows:

A. Cluster Analysis Model

Clustering algorithm is first deployed to group warehouses with similar characteristics. Distance-based clustering algorithm such as K-means clustering, or X-means clustering [11, 12], is used to establish the number of clusters to separate groups of data points and to maximize the cluster separation between close data points.

Other clustering algorithms such as Expectation Maximization (EM) clustering can also be used [11]. The clustering algorithms' performance are measured by analyzing the internal cluster similarity (using statistical measurements such as mean, median and standard deviation).

Good cluster separation is achieved when the individual features can differentiate the characteristics of cluster of warehouses separated by the clustering algorithm. Rules which describe each cluster of warehouses can be derived using a Decision Tree Algorithm or Associative Rule Mining, with the individual features as the input, and the cluster assignment set as the predictor.

B. Linear Regression Analysis

Linear regression analysis are able to identify individual dependent variables that have a strong correlation with the independent variable, for example, energy consumption. Dependent variables with a high degree of correlation with energy consumption such as area of warehouse, and the energy consumption by cooling/heating degree days were identify using the linear regression analysis.

C. Decision Tree Model

For each warehouse sample, the cluster assignments derived from the clustering exercise on the Cluster Groups, along with other variables describing the warehouse building and operational characteristics, are then used to train a decision tree model, which predicts the expected energy consumption of warehouses. The summarized steps to train the decision tree are as follows:

1. For each warehouse example, the clusters assigned by the Cluster Analysis Model are used as input variables. Each cluster assignment variable encodes summarized information about the warehouse characteristics.
2. The energy consumption variable derived from the Linear Regression Analysis is set as the predicted variable.
3. The predicted variable is binned (or discretized) into equal sized bands. The energy consumption band assigned to each warehouse example will be the Decision Tree's predicted variable.
4. The data set is split into a training and test set. The training set (typically 70% of the data set) is used to train the decision tree model to predict the correct energy consumption band. The quality of the decision tree predictions is measured by the accuracy of the decision tree model predicting the remaining (typically 30% of data set) data.

Further experimentation on binning distributions can also be conducted to find a binning distribution, which works best for the accuracy and complexity of the decision tree model. Since the decision tree model is trained to predict the likely energy consumption band, fewer and wider bins will result in higher accuracy in prediction as the widths of the bins are larger, but will result in less differentiated groups of warehouses.

The experiments are conducted by making necessary adjustments variables and parameters as described below.

- Adjusting the variables and parameters used to derive the Cluster Analysis Model and Linear Regression Analysis, such as the number of clusters in each cluster group in the Cluster Analysis Model, and re-examining the correlation of the variables in the Linear Regression Analysis,
- Adjusting the number and the width of bins of the predictor variable in the Decision Tree model,
- Changing the algorithm parameters of the Decision Tree model, for example the minimum support and minimum number of warehouse examples in each decision tree node, and the decision tree complexity (tree depth).

V. FRAMEWORK, ANALYSIS AND DATA-DRIVEN EXPERIMENTS

We proposed a systematic framework for conducting the study. The framework is summarized in Figure 2.

In the initial data analysis steps, questionnaire with 10 distinct categories of questions relating to warehouse building

and operational characteristics (see Figure 2) are to be given to warehouses for data collection. In our study, we are using the questionnaire data from U.S. Energy Information Administration (EIA) as our initial step to collect the data.

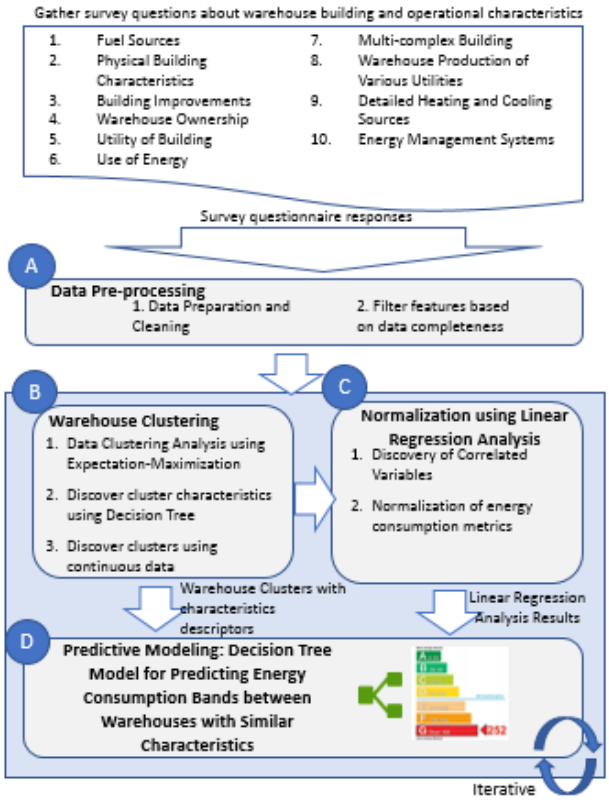

Figure 2. Systematic Framework of Data Clustering and Decision Trees to Predict Energy Consumption band

A. *Pre-processing data*

Pre-processing the data before training the predictive model simplifies the interpretation of the results, and help us extract meaningful insights. The first data preparation work involves basic data cleaning for data completeness. Next, we pre-process the data by summarizing the continuous variables into discrete "binned" variables, or binary ("Yes or No") responses. Decision were made for missing responses to be either considered a separate "Missing" classification, or to be left out of the analysis if there are too many missing features.

B. *Warehouse Clustering*

As discussed previously, warehouse energy consumption benchmark is more meaningful if it is benchmarked against warehouses of similar characteristics. Therefore, the first task in our analysis is to perform data clustering analysis to group warehouses based on their similarities. To do this, we ran clustering analysis on each of the 10 groups of questions. X-means clustering [9] is used to estimate the number of suitable clusters, and derive cluster characteristics. It is based on the k-means clustering principle and works on small data set. In addition, 81 binary questions ("Yes" / "No" responses) are binned into 10 cluster groups by their context and granularity of the responses. The Expectation-Maximization (EM) algorithm is applied to the clustering analysis. Within each cluster analysis, a cluster is assigned to the warehouse.

A combination of 33 continuous and categorical variables were used as input features to cluster 473 samples of warehouses. The Expectation Maximization algorithm is set to a maximum iteration of 100 and 100 seeds. In applying the Expectation Maximization algorithm, we set:

$D = \{x^{(1)}, \ldots, x^{(n)}\}$ be n observed data vectors (features of the data set).

$Z = \{z^{(1)}, \ldots, z^{(n)}\}$ be n values of hidden variables (i.e., the cluster labels).

The heuristics function was used to determine an optimal number of 6 clusters {0,…, 5}. The cluster frequency is shown in Figure 3. The clustering results converge, as repeated runs of the clustering algorithm did not change the cluster distribution.

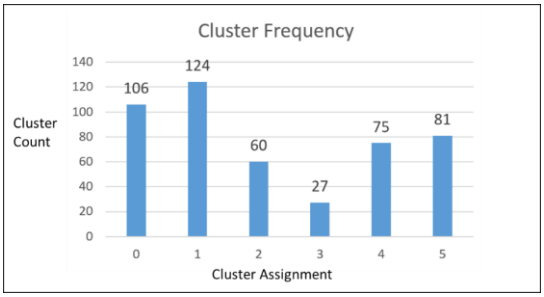

Figure 3. Cluster Frequency of 33 Input Variables

A decision tree model can be trained with the cluster assignment as the predicted variable to uncover the cluster characteristics. The cluster is then set as the predicted variable. A decision tree using Bayesian Dirichlet Equivalent with Uniform Prior (BDEU) [10] as it's splitting score is modeled to uncover characteristics of each cluster of warehouses. Two different experiments were conducted to obtain the decision tree rules.

Experiment 1. Complex Decision Tree Model to Recover Cluster Characteristics

The first experiment creates a complex decision tree structure that will correctly classify the maximum number of examples. This often results in an over-fit decision tree as it attempts to characterize all the examples in the data set down to a single record (in theory, the minimum support of the decision tree rule would be 1/sample size). Such a decision tree maximizes the accuracy at the expense of model complexity.

The complex decision tree has a minimum support of 1 case. The model is set at 15% holdout rate for testing, where 15% of data set samples were left out from training the decision tree, and later re-introduced to score the model's accuracy. 403 examples were used to train the decision tree model while 70 examples were used to score the model's accuracy. The model classifies 95% of test example cases to the correct cluster.

The rules which describe each cluster (C0, …, C5) is summarized below. Decision tree nodes with fewer than 5 cases were omitted (WH – denotes Warehouses, A/C – denotes Air-conditioning).

- C0: WH with no heating → < 10% cooled → < 100,000 sq feet
- C1: WH using Individual Space Heaters → >90 % Cooled by Package heaters.
- C1: WH with no heating → > 10% cooled → > 10% by Package A/C
- C1: WH using Boilers or Package Heating → > 50% Cooled by Package A/C
- C1: WH using district steam/hot water or other heating equipment → > 10% Cooled by Package A/C
- C2: WH using hot air from furnace → > 90% Cooled by Package A/C
- C3: WH using hot air from furnace → < 90% Cooled by Package A/C → > 60% Cooled by Heat Pumps
- C3: WH using heat pumps → > 80% Cooled by Heat Pumps → Not Equal Glass on All Sides
- C4: WH using hot air from furnace → < 90% Cooled by Package A/C → < 60% Cooled by Heat Pumps
- C5: WH using Individual Space Heaters → < 10% Cooled by Package Heaters.
- C5: WH using Boilers or Package Heating → < 50% Cooled by Package A/C → 10% Lit when Closed
- C5: WH using district steam / hot water or other heating equipment → < 10% Cooled by Package A/C

Experiment 2. Generalized Decision Tree Model to Recover Cluster Characteristics

The second experiment is designed to test a decision tree model, which creates a more generalized set of rules that are easy to understand. This implies that there will be fewer sets of rules, which describe the characteristics of each cluster of warehouses. The decision tree would have an acceptably lower accuracy due to the fewer and simpler rules, but each rule generated will have a higher support, for example minimum support of 2.5% of sample size.

The generalized decision tree is has a minimum support of 2.5%, or 10 cases. In this instance, model is set at 30% holdout rate for testing. The model achieves an 89% accuracy in predicting the right cluster in 141 test examples.

The following summarizes of the rules which describe each cluster (C0, … , C5).

- C0: WH with no heating → < 10,000 sq feet → No auto-lighting installed
- C1: WH using boilers / Package Heating / Individual Space Heaters → > 90 % Cooled by Package coolers.
- C2: WH using hot air from furnace → > 90% Cooled by Package A/C
- C3: WH using heat pumps / district steam or hot water / other heating source
- C4: WH using hot air from furnace → < 90% Cooled by Package A/C
- C5: WH using boilers / Package Heating / Individual Space Heaters → < 90 % Cooled by Package coolers.

The decision tree shows warehouses in Cluster 0 are generally those of small areas, and no heating. Other clusters are influenced by the type of heating equipment, and amount of required cooling by equipment type.

The results from the complex and generalized variants of the decision tree model allows important features about groups of warehouses to be uncovered. The unique properties about each cluster can be used to group warehouses to evaluate their energy consumption among those groups. Depending on the complexity of the decision tree rules, more clusters of similar warehouses can then be derived from more complex sets of rules.

Discover Additional Clusters from Continuous Energy Consumption Data

Up to this point, input features about a warehouse's building characteristics are clustered to simplify the process of finding similar groups of warehouses. In this step, we can take energy consumption patterns of individual warehouses to find groups of warehouses with similar consumption characteristics. The consumption characteristics encode information about the warehouse's energy-related operational characteristics. Different fuel sources are used for different purposes. Fuel sources may be used for lighting, heating/cooling, heavy or electric machinery, computers and servers, etc. The types of fuel source will vary, and warehouses typically set-up their heating needs with specific sources of energy, for example Natural Gas and District Heating sources. Analyzing the warehouse energy consumption patterns is a useful step in determining the efficiency rating of a warehouse.

Cluster Assignment	Electricity used	Natural Gas used	Fuel oil / diesel / kerosene used	Bottled gas / LPG /propane used	District steam used	District hot water used
cluster0	79%	0%	6%	12%	0%	1%
cluster1	100%	100%	0%	9%	1%	0%
cluster2	100%	83%	100%	40%	0%	0%
Total	89%	50%	9%	12%	1%	0%

Figure 4. Cluster Group 1 - Cluster Characteristics of Fuel Source used

Figure 4 shows the proportion use of fuel sources of warehouses examples in each cluster. While this cluster analysis is useful to identify the fuel sources, it does not provide sufficient information about the proportion of actual energy consumption attributed to each fuel source. Hence, we can utilize the energy consumption for each fuel source (normalized to BTU-equivalent units), and find clusters from the degree/proportion attributed to each fuel source as a better indicator of attainable warehouse energy efficiency. In the buildings data set, a building can derive its energy consumption from a variety of fuel sources. Energy fuel sources include Electricity, District Heating source, Fuel Oil, Natural Gas. We can approximate the expected energy consumption efficiency by the proportion of energy consumption attributed to the various fuel sources.

Expectation Maximization (EM) clustering algorithm is then run on the converted variables for every warehouse, to

identify clusters of warehouses with similar fuel sources, and proportions of fuel sources.

Figure 5 shows a simplified interpretation of the results. Each cluster has a distinct set of fuel sources, and the cluster analysis is able to identify warehouses with significant sets of warehouses according to their similarity.

Cluster Assignment	Cluster Proportion of Entire Data set	Electricity used	Natural Gas used	Fuel oil / diesel / kerosene used	District Heat
cluster0	51%	100%	18%	0%	0%
cluster1	1%	100%	20%	0%	100%
cluster2	4%	100%	32%	100%	0%
cluster3	43%	100%	100%	0%	1%
cluster4	17%	93%	100%	1%	0%
Total	100%	99%	54%	5%	1%

Figure 5. Cluster Characteristics of Consumption Proportions

We can interpret the results of each cluster. Table I shows a sample of the proportion of consumption from each fuel source of each warehouse example in cluster1. The cluster is characterized as warehouses with 30% energy sourced from electricity, and 70% from district heat.

TABLE I. SAMPLES OF CLUSTER1 – FUEL SOURCE CONSUMPTION PROPORTIONS OF EACH WAREHOUSE EXAMPLE

Warehouse Example	Electricity	Natural Gas	Fuel Oil	District Heat
1	0.2	0.2	0.0	0.6
2	0.3	0.0	0.0	0.7
3	0.3	0.0	0.0	0.7
4	0.4	0.0	0.0	0.6
5	0.3	0.0	0.0	0.7

With this analysis, further characteristics of the warehouse clusters were identified.

C. *Normalization using Linear Regression Analysis*

We noticed, Heating and Cooling Degree Days (HDD and CDD) are two variables affecting the energy consumption of warehouse, as heating and cooling demand for energy to heat or cool a building. HDD and CDD are influenced by measurements of outside air temperature. Figure 6 shows the variation in demand of energy in terms of heating and cooling degree days, deviation from a base 65 degrees Fahrenheit or 18 degrees Celsius.

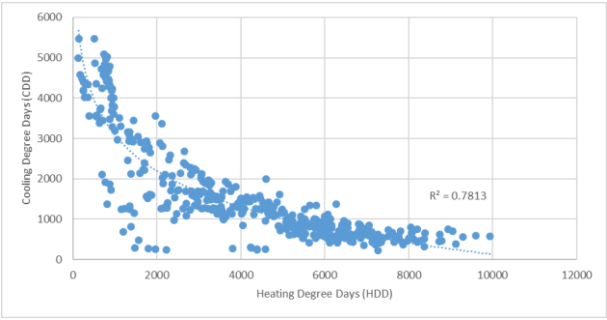

Figure 6. Variation in relationship between Heating Degree Days and Cooling Degree Days

The heating requirements for a given structure at a specific location are directly proportional to the number of HDD or CDD at that location. The heating and degree days are inversely proportional to each other due to the climate zones. For example, buildings with a very high HDD is unlikely to require much cooling as it is likely situated in a warm climate, and likewise buildings with a very high CDD would not require much heating. Hence, the climatic differences alone can affect the energy demands, and the type of heating/cooling equipment used.

Finding Linear Relationship to Energy Consumption based on Floor Area

The dependent variable in this part of the study is the Major Fuel Consumption (MFBTU8) variable, an energy consumption number combining electricity, natural gas, and fuel oil sources.

The following information in Table II, adapted from the U.S. Energy Information Administration (EIA) describes how the different energy consumption components are converted from source unit to BTU Equivalent, and combined in MFBTU8 (Annual major fuel consumption in thousands of BTU).

TABLE II. U.S. ENERGY INFORMATION ADMINISTRATION (EIA) CONVERSION OF ENERGY SOURCE TO BTU EQUIVALENT UNITS

Energy Source	BTU Equivalent	Unit
Electricity	3,412	kilowatt-hour
Natural Gas (2003)	1,031	cubic Foot
Distillate Fuel Oils (Nos. 1, 2, and 4)	138,690	Gallon
Residual Fuel Oils (Nos. 5 and 6)	149,690	Gallon

The floor area directly influences energy consumption. The larger the gross floor area warehouse building, the likely increase lighting, heating and cooling requirements. The type of bulb correlates to the energy consumption band. The high use of incandescent light bulbs is likely to result in higher energy consumption compared to fluorescent light bulbs. This is due to the higher energy efficiency of fluorescent and LED light bulbs compared to conventional incandescent bulbs (Table III).

TABLE III. GUIDE TO MORE EFFICIENT AND MONEY-SAVING LIGHT BULB. NATURAL RESOURCES DEFENSE COUNCIL.

Target Brightness (Lumens)	Standard Incandescent (Watts)	Halogen Bulbs (Watts)	Fluorescent (CFLs) (Watts)	(LEDs) (Watts)
450	40	29	9	8
800	60	43	14	13
1100	75	53	19	17
1600	100	72	23	N.A. currently

Figure 7 shows the relationship of floor area in square feet of each warehouse building, in relationship to its annual total energy consumption in BTU-equivalent units (MFBTU8 variable in the data set).

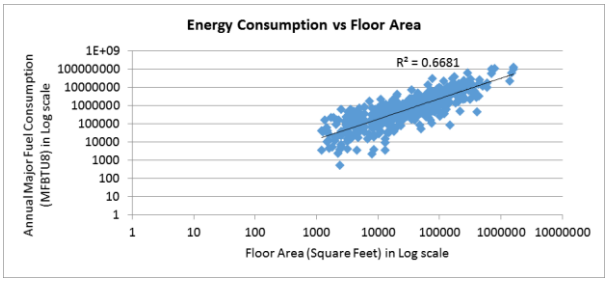

Figure 7. Energy Consumption in relation to Floor Area

Finding Linear Relationship to Energy Consumption based on Work Hours

The working hours each week will affect the activity of the warehouse site. Intuitively, the number of work hours multiplied by the floor area would be a better approximation of the energy consumption. Figure 8 shows the stronger relationship between annual energy consumption of warehouse buildings versus Floor Area multiplied by average weekly work hours.

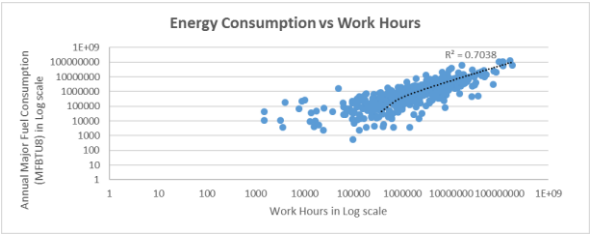

Figure 8. Energy Consumption in relation to (Floor Area * Work Hours)

Finding Linear Relationship to Energy Consumption based on Floor Area, Work Hours, and CDD/HDD

Combining the previous three relationships (Floor Area, Work Hours, Cooling/Heating Degree days) with respect to energy consumption, a combined metric can be derived to show a normalized energy consumption figure. As mentioned previously, energy consumption is related to the demand for energy in different climate zones, because of heating and cooling requirements. Therefore, a better representation of energy consumption could be *Energy Major Fuel Consumption per Heating or Cooling degree days*. Assuming there are negligible differences in the efficiency required to heat or cool a building per degrees Celsius/Fahrenheit, we can normalize the predicted energy consumption as

$$\frac{Energy\ Consumption}{(CDD + HDD)}$$

and compare it against (Work Hours * Floor Area). Figure 9 shows a strong R-square = 0.7428 measure for correlation between the two combined metrics, and implies the three variables (Floor Area, Work Hours, Cooling/Heating Degree days) have a strong influence on the prediction of energy consumption. Hence, the combined metric

$$\frac{Energy\ Consumption}{(CDD + HDD)}$$

can be chosen as the predicted variable as an alternative to the raw BTU-equivalent energy consumption metric, later in the decision tree model.

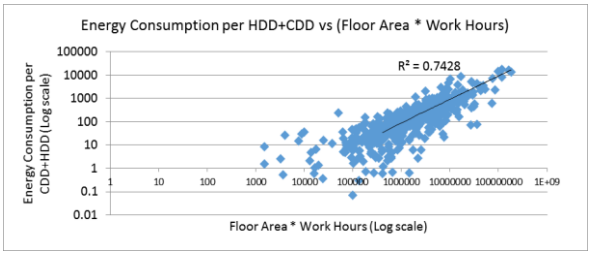

Figure 9. Relationship between Energy Consumption per HDD+CDD and Floor Area, Work Hours

D. Predictive Modeling: Decision Tree for Predicting Energy Consumption Band (for benchmarking)

A decision tree using Bayesian Dirichlet Equivalent with Uniform Prior (BDEU) as its splitting score is modeled to predict the Energy Consumption bands of warehouses with similar characteristics. From the decision tree nodes, we can examine the properties of warehouses with similar energy consumption and similar building and activity characteristics, since the Cluster Groups, and other continuous and categorical variables will be encoded as decision nodes.

All warehouse examples are first ranked from the lowest to highest energy consumption. They are then separated into bins of approximately equal number of warehouses, and the continuous energy consumption values are discretized into five bands accordingly. Table IV describes the 5 individual energy consumption bands with approximately equal frequency distribution.

TABLE IV. WAREHOUSE ENERGY CONSUMPTION PREDICTION BY DISCRETISED BINS

Energy Consumption Band	MFBTU8 (in thousands of BTU)	MFBTU8 / (HDD+CDD) [Derived Consumption Metric]	Number of Cases	Prob-ability
1	<101,084	< 19.724	84	19.9%
2	101,084 – 346,147	19.724 - 64.575	84	19.9%
3	346,147 – 974,473	64.575 - 168.438	85	20.1%
4	974,473 – 3,211,618	168.438- 578.556	84	19.9%
5	>= 3,211,618	>= 578.556	86	20.3%
Count			423	100.0%

Two more experiments were conducted to compare the accuracy of predicting the warehouse energy consumption bands. A third experiment was conducted to train the decision tree prediction model with only the continuous and categorical features derived from Part B (Warehouse clustering) to predict energy consumption. The decision tree structure resulted has a depth of up to 9 levels, most of which are binary decision splits. This makes interpretation of results challenging as it is difficult to gain useful business rules about similar warehouses within each energy consumption band.

A fourth experiment was conducted to train the decision tree prediction model by combining the results from Part B and Part C, we included the

- Warehouse Cluster Groups,
- Fuel Sources Consumption Cluster (warehouse clustered according to proportion of various fuel sources electricity, natural gas, fuel oils, and district heat),
- Derived metrics from the Linear Regression analysis (Work Hours * Floor Area),
- Derived Consumption Metric $\frac{\text{Energy Consumption}}{(\text{CDD} + \text{HDD})}$ as opposed to the raw consumption metrics,
- Additional Continuous and Categorical variables.

The decision tree for both experiments are trained with 70% of the warehouse examples. The accuracy of the decision tree model to predict warehouse energy consumption is the proportion of correctly identified discrete energy consumption band in the remaining 30% (test) warehouse examples. The minimum support (number of cases) is set at 2.5%.

The fourth Experiment yielded an improved overall accuracy of 52% in predicting test cases with 5 levels of depth in the Decision Tree. We therefore concluded that combining Warehouse Clustering methods and Linear Regression Analysis yielded better results with the given data set.

VI. Discussion

With the data-driven methodologies described in this paper, decision makers can cross-validate the energy rating provided by their existing models and make changes to their existing assumptions and methods to improve their benchmarking results. Our work complements the existing statistical, analytical, and simulation methods by attempting to provide a way to interpret the results of existing benchmark models. The Cluster Analysis model provides insights into the similarity and differences between groups of warehouses, while the Linear Regression Analysis provided additional parameters and metrics to improve the energy band (energy benchmark) prediction.

The following strategy matrix provides a summary of techniques that can be employed at each phase of the energy efficiency benchmark study.

Task vs Techniques	Clustering	Decision Tree	Linear Regression
Discovering Warehouses with Similar Characteristics	Cluster multiple variables to find cluster groups	Derive cluster characteristics from cluster groups.	
Finding various factors affecting Energy Consumption.		Discretize continuous variables and train a decision tree to predict energy consumption	Conduct linear regression to find correlated variables.
Comparing Energy Consumption among Similar Warehouses	Cluster variables to find similar warehouse characteristics	Build a decision tree with discrete and continuous input variables	Use Linear Regression results for improving result

VII. Conclusion

The paper described an iterative data analysis and data-driven method to energy certificate benchmarking. Multiple data-mining algorithms applied at the various stages of data analysis help decision makers uncover new insights to warehouses' building, operational activity levels, and energy demand requirements because of weather climate differences. We showed that through the use of clustering and decision tree model, we are able to provide descriptive rules about similarity and differences of warehouses that make benchmarking study results more apparent, comparable and practical for stakeholders. Decision-makers can better adopt relevant and impactful best practices to improve the energy consumption of their warehouses based on actionable insights from the models. Our generalized (non-expert opinionated) process is practical and customizable, making it adaptable and applicable to ever-changing requirements, energy reduction technologies, energy sources, and environmental policies.

References

[1] A. McKinon, M. Browne, A. Whiteing, M. Piecyk, W.-K. Chen, *Green Logistics: Improving the Environmental Sustainability of Logistics*, 2015.

[2] E. Burman, S.M. Hong, G. Paterson, J. Kimpian, D. Mumovic, "A comparative study of benchmarking approaches for non-domestic buildings: Part 2 – Bottom-up approach", *International Journal of Systainable Built Environment*, vol. 3(2), pp. 247-261, Dec 2014.

[3] M. F. Keohane, "Energy Benchmarking for Commercial Buildings", *Sustainability Energy Authority of Ireland*, 2013

[4] R. Liddiard, A. Wright, L. Marjanovic-Halburd, "A review of non-domestic energy benchmarks and benchmarking methodologies", *In Proceedings of the IEECB Focus 2008 Improving Energy Efficiency in Commercial Buildings Conference*, vol. 3(2), Congress Center Messe Frankfurt (CMF), Frankfurt am Main, Germany, April 2008.

[5] L.F. McGinnis, W. Chen, P. Griffin, G. Sharp, T. Govindaraj, D. Bodner, "Benchmarking Warehouse Performance", *School of Industrial & Systems Engineering Georgia Institute of Technology Atlanta*, 2002.

[6] Portela, M.C.A.S, E. Thanassoulis, "Developing a decomposable measure of profit efficiency using DEA," *Journal of the Operational Research Society*, vol. 58(4), pp. 481–490, 2007.

[7] N. N. Abu Bakar, M. Y. Hassan, H. Abdullah, H. A. Rahman, M. P. Abdullah, F. Hussin, M. Bandi, "Energy efficiency index as an indicator for measuring building energy performance: A review", *Renewable and Sustainable Energy Reviews*, vol. 44, pp 1-11, April 2015.

[8] P. Jiang, Y.H. Chen, W.B. Dong, B.J. Huang, "Promoting low carbon sustainability through benchmarking the energy performance in public buildings in China", *Urban Climate*, vol. 10(1), pp. 92-104, Dec 2014.

[9] EPA Technical Reference, "Energy Star Scores for Warehouses in the United States". *Energy Star*, Nov 2014.

[10] EPA Data Trends, "Energy Use in Non-Refrigerated Warehouses", *Energy Star*, Jan 2015.

[11] D. Pelleg, A. Moore, "X-Means: Extending K-means with Efficient Estimation of the Number of Clusters", *International Conference on Machine Learning*, 2000.

[12] N. Alldrin, A. Smith, D. Turnbull, 2003. "Clustering with EM and K-Means," *University of San Diego, California, Tech Report*, 2003.

[13] D. Heckerman, D. Geiger, and M. Chickering, "Learning Bayesian networks: the combination of knowledge and statistical data", *Machine Learning*, 1995.