Use of Machine Learning Algorithms to Propose a New Methodology to Conduct, Critique and Validate Urban Scale Building Energy Modeling

by

Maharshi Pathak

A Thesis Presented in Partial Fulfilment
of the Requirements for the Degree
Master of Science

Approved July 2017 by the
Graduate Supervisory Committee:

T Agami Reddy, Chair
Marlin Addison,
Harvey Bryan

ARIZONA STATE UNIVERSITY
August 2017

ABSTRACT

City administrators and real-estate developers have been setting up rather aggressive energy efficiency targets. This, in turn, has led the building science research groups across the globe to focus on urban scale building performance studies and level of abstraction associated with the simulations of the same. The increasing maturity of the stakeholders towards energy efficiency and creating comfortable working environment has led researchers to develop methodologies and tools for addressing the policy driven interventions whether it's urban level energy systems, buildings' operational optimization or retrofit guidelines. Typically, these large-scale simulations are carried out by grouping buildings based on their design similarities i.e. standardization of the buildings. Such an approach does not necessarily lead to potential working inputs which can make decision-making effective. To address this, a novel approach is proposed in the present study.

The principle objective of this study is to propose, to define and evaluate the methodology to utilize machine learning algorithms in defining representative building archetypes for the Stock-level Building Energy Modeling (SBEM) which are based on operational parameter database. The study uses "Phoenix- climate" based CBECS-2012 survey microdata for analysis and validation.

Using the database, parameter correlations are studied to understand the relation between input parameters and the energy performance. Contrary to precedence, the study establishes that the energy performance is better explained by the non-linear models.

The non-linear behavior is explained by advanced learning algorithms. Based on these algorithms, the buildings at study are grouped into meaningful clusters. The cluster

“mediod” (statistically the centroid, meaning building that can be represented as the centroid of the cluster) are established statistically to identify the level of abstraction that is acceptable for the whole building energy simulations and post that the retrofit decision-making. Further, the methodology is validated by conducting Monte-Carlo simulations on 13 key input simulation parameters. The sensitivity analysis of these 13 parameters is utilized to identify the optimum retrofits.

From the sample analysis, the envelope parameters are found to be more sensitive towards the EUI of the building and thus retrofit packages should also be directed to maximize the energy usage reduction.

`

DEDICATION

In fond memory of my grandmother, naniji,

Who did not live to see this complete.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my committee chair, Dr. T Agami Reddy for the continuous support during my Master's study and project, for his patience, motivation, and vast knowledge of statistical design experiments. His guidance helped me in forming and fine tuning the research and proposed methodology of this project.

I would also like to thank my committee member, Marlin Addison, who have been teaching me for past two years the art and science of building energy performance studies and his vocabulary sessions every Thursday.

Muthu Ramalingam has been a source of inspiration since the start of my MSBE studies. His office hours have been a flash course of its own kind.

Special thanks to program coordinator and committee member Dr Harvey Bryan, who supported and facilitated my studies in every possible instance.

My family members especially my father who have been supportive and encouraging at all times. His belief in my capabilities have pushed me to set tougher standards for myself. The two lovely ladies of my beloved family, my mother and sister, thanks for reminding me to always be humble and grounded and not let my guard down. Your endless love and support was my motivation.

Last but not the least, the crazy bunch of friends one can ever imagine, Shriya, Meha, Meenal and Aviral, whose support in the last stage of my thesis was key to completing the tasks at hand. Thanks for making all those all-nighters possible.

TABLE OF CONTENTS

Page

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

1. INTRODUCTION 1

1.1 Background 1

1.2 Scope and Objectives 2

2. LITERATURE REVIEW 3

2.1 Introduction 3

2.2 Review of Studies focused on “Urban Energy Modeling” 3

2.2.1Building Energy part of the Study Focus 5

2.2.2Building Energy Performance – The Primary Study Focus 5

2.3 Urban Building Energy Models 9

2.3.1 Bottom-up Approach 9

3. METHODOLOGY & THEORY 12

3.1 Introduction 12

3.2 Experimental Design 15

3.2.1 Exploratory Data Analysis (EDA) 16

CHAPTER Page

3.3 Random Forest 17

3.3.1 Random Forest Algorithm 18

3.3.2 Out of Bag Observations and forecasting error 19

3.3.3 Predicator Importance with RF 20

3.3.4 RF Tuning Parameters 21

3.4 Clustering 22

3.4.1 Agglomerative Hierarchical Clustering 24

3.4.2 Clustering Linkage 26

3.5 Building Energy Simulation 30

3.5.1 Theory 30

3.5.2 Sample Building Description 32

3.6 Post-Processing and Validation 35

3.6.1 Monte-Carlo method 35

3.6.2 Latin Hypercube Sampling 39

4. ANALYSIS AND DISCUSSION 41

4.1 CBECS Database 41

4.2 Data Preprocessing 42

4.2.1 Data Filtration 42

CHAPTER Page

4.2.2 Selection of Variables and Multi Linear Regression Analysis 46

4.2.3 Results of OLS Model of CBECS Data 47

4.2.4 Selection of building with the climate conditions like Phoenix 50

4.2.5 Random Forest 51

4.2.6 Hierarchical Clustering 56

4.3 Building Energy Simulation Analysis 62

5. RESULTS AND CONCLUSION 72

REFERENCES 77

APPENDIX

A: EDA OF THE CBECS SELECT VARIABLES 79

B:PNNL PROTOTYPE MEDIUM SIZE OFFICE BUILDING 86

LIST OF TABLES

Table Page

1: Building Parameters Used for The Uncertainty Analysis..... 35

2: Summary of Data Filters Considered for The Screening Criteria 43

3: Cbecs Database – Variable Selected for Regression 46

4: Cluster Mediods after Conducting the Hierarchical Clustering on Select CBECS Data Set -PhX Dataset..... 61

5: Variable Details of Cluster Selected for Method Validation and Monte- Carlo Simulation..... 61

6 Sensitivity Analysis of the Variables on the Batch Simulation Results - Pearson Coefficient Value for Each Variable..... 67

7 Descriptive Analytics of the Selected Important Variables of Cluster1 74

LIST OF FIGURES

Figure Page

1: Distribution of Publication Dates of Papers Matching the Topic (Urban or City) Energy Model 4

2: Type of Building Performance Evaluation Models 7

3: MLPDM Flowchart: 13

4: Principle of Statistical Analysis –..... 16

5 Agglomerative/Divisive Hierarchical Cluster..... 23

6 Agglomerative Hierarchical Clustering Algorithm Flowchart 25

7 Lance and Williams' Parameters for Agglomerative Hierarchical Clustering. 27

8: Simulation Strategy - Real Building to Model Calibration 32

9: TMY2 Weather Data: Phoenix 33

10: PNNL Reference Prototype Office Building Geometry - Courtesy: PNNL Scorecard 34

11: Zoning Pattern of the Reference PNNL Office Building Prototype - Courtesy: Pnnl Score Card..... 34

12 Latin Hyper Cube Sampling- Two Dimensional 40

13: Gross Area Distribution..... 44

14: Histogram for Cdd/Hdd 45

15: Primary EUI Distribution..... 45

16: Ols Model Results..... 47

17: Correlation Matrix Based On Pearson Coefficient..... 48

Figure Page

18 Random Forest Algorithm R Code 51

19 Variability Explained by the RF Algorithm..... 52

20 CBECS Data vs RF Model Prediction 52

21 : Error vs Number of Trees- RF 53

22 Tuning the Random Forest Algorithm 54

23 Rf- Variable Importance Plot 55

24: Hierarchical Clustering:Dendrogram- Select CBECS Data 57

25 9 Clusters and Their EUI Distribution Range Comparison 58

26: Variable Distribution Plot For The Clusters: CBECS PhX Climate 59

27: Discriminant Plot for Each Cluster 60

28: The Building Geometry of Cluster 1 62

29: The Reference vs Modeled Eui 64

30: Building #1 EUI Histogram -Monte Carlo Simulation 66

31 Parallel Coordinate Plot 68

32 5 Varibles with Positive Correlation with EUI 69

33 Comparing the Eui Distribution for the Select 5 to Derive Better EUI Targets. 70

34 Most Effective Combination of the Envelope Improvement Strategies with Effective Data Visualization Technique and What-If Analysis 71

35 Clusters and Their % EUI Distribution Compared to the Eui Distribution of the Whole Data Set 73

36 Eui Distributions of Each Building's 1000 Simulations 75

37 Eui Distributions of the Representative Mediod Building's 1000 Simulations:. 75

38: Select Design Variables' Interaction with the Response Variable - EUI 81

39: Select Operational Variables (-1) ' Interaction With EUI..... 82

40: Select Operational Variables (-2) ' Interaction With EUI..... 82

41 Design Variables - Correlation by Data Mining..... 83

42 Operational Variable- Plug Load - Correlation Matrix..... 84

43 Operational Variable-Occupancy Load - Correlation Matrix..... 84

1. INTRODUCTION

1.1 Background

Urban scale building performance analysis has emerged as a multi nodal multi-criterion (MNMC) optimization exercise. This exercise can enable the researchers to understand and model observed energy consumption patterns and predict the future behavior based on these patterns. MNMC optimization assumes that each node would have a characteristic building associated with it and each building's complexity originates from the large number of variables involved, from the dynamic nature of building loads and processes, from the intricacy of interaction effects among variables, and from the inability of the research team to view cause and effect in multi-dimensional space. Around 3000 input variables are required when a building is considered for whole building energy simulation on a simulation engine such as ENERGYPLUS (Crawley, et.al. '99). Conducting the same for a large number of buildings makes the problem highly complex and beyond human intuition. Modern day statistical advancements allow users to address this large-scale data gathering, exploration and analysis feasible. Based on these machine learning algorithms, on information based automated methodology needs to be developed which can act as a bridge between the whole building energy simulation engine and the statistical analysis software.

This study proposes a semi-automated methodology to create and validate a novel building clustering technique which would enable stakeholders make informed decisions towards improving energy consumption reduction targets for the proposed study area, assess impact of potential retrofits on a larger scale, to understand existing energy supply and consumption patterns and to obtain newer supply alternatives.

1.2 Scope and Objectives

The research aims to propose, evaluate and validate a new methodology to create prototypical buildings used in urban scale building energy modeling based on knowledge obtained from the realm of Big data analytics specifically.

a) The study aims to identify building energy performance indicators based on machine learning regression algorithms under the hypothesis that the relationship between performance indicators and response variable should a non-linear relationship.
b) Further, identify clustering algorithms for dividing the large data base of the study area into meaningful clusters under the hypothesis that the “mediod” of the cluster, created based on key performance indicator can be representative of the buildings contained in a cluster.

2. LITERATURE REVIEW

2.1 Introduction

Building energy modeling of existing buildings involve uncertainty and sensitivity associated with the parameters under study due to various factors. These include insufficient details required for defining the building parameters, discrepancy between model vis-a-viz true behavior of existing building or manual errors. The thesis study aims to focus on identifying impact due to uncertainties of the specific parameters, and solving the discrepancies between model and actual building behavior on an urban scale building energy usage patterns by validating the models with measured data on varied temporal scale (i.e. annual, monthly, daily, hourly).

2.2 Review of Studies focused on "Urban Energy Modeling"

Urban/city scale energy modeling has been of keen interest for researcher in very recent times. As Reinhart et al. puts in their study [Reinhart et.al., 2015], it's a nascent field and novel approaches are being studied extensively. A detailed review (number of articles reviewed by the study are sorted in form of their date of publication in FIGURE) in the field of UEM has been performed by James Keirstead in his review notes studies of urban energy systems can be attributed in following ways.

1) Temporal and spatial

2) Methodology

3) Appliances and target audience

4) Supply and demand

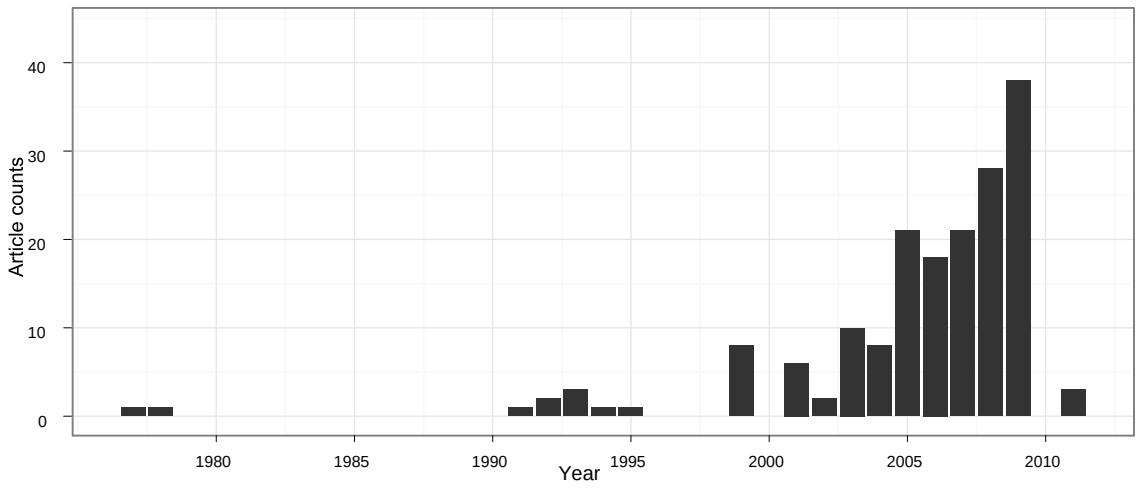

FIGURE 1: DISTRIBUTION OF PUBLICATION DATES OF PAPERS MATCHING THE TOPIC (URBAN OR CITY) ENERGY MODEL [1]

Present study aims to focus on the demand side of the urban energy modeling. This can be further classified into dealing with building design and renovation, energy demand estimation in the built environment, urban climate as it directly affects buildings, urban planning and policy, and transport. They represent a range of spatial scales, from single buildings to groups of buildings in a street or district or the whole city, and the behavior of individuals. Temporal scale is also varied, with the three most common scales being static, annual time-series or hourly [Kierstad et.al., 2012]. Amongst these broad classifications, the present study focuses on identifying building design, ways to characterize it and estimating the energy demand based on these criteria and analyzing optimum retrofit packages amongst available options.

In the last decade, few studies have been conducted focusing on Urban/ block scale building energy modeling. These studies can be broadly divided in two types based on their approaches towards achieving the goal.

2.2.1Building Energy part of the Study Focus

Studies under this type of scale are generally top-down (mathematical/ statistics based) models and are of large radius which covers the study area. Study focused on urban microclimates and its impact on energy use are of focus of this studies, building energy use are just a part of the whole study and not the focus. These types of studies include impacts of meteorological changes via WRF (weather research and forecasting) models, canyon effects, microclimate models [Salamanca et.al. 2014, 2015, 2016, Chen et.al., 2011, Dorer et.al. 2013, Ozkeresteci et.al. 2003]. So, the results of these are not of a much help when making policy decisions specific towards efficient energy use in the building sector.

2.2.2Building Energy Performance – The Primary Study Focus

To understand the urban scale modeling studies conducted, first let's understand how individual buildings thermal performance is modeled and analyzed in the context of existing buildings. Coakley et.al. (2014) provides a detailed review regarding these studies. According to this review paper, the studies with focus on building energy performance as focal study point have 2 types of approaches towards studying the thermal behavior of the buildings (FIGURE 2).

i. Law driven or forward model: models driven by laws of physics such as mass/heat transfer phenomenon. Models based on these laws provide detailed explanation and

reasoning behind the working of the system which are not captured by behavioral prediction models.

ii. Data driven or Inverse model: Inverse model works with the behavior of the systems and derives methods to describe the systems via mathematical equations and regression models. For this reason, behavioral models of large scale can be understood with minimum number of variable inputs. The data driven or inverse models can be further divided into 3 major types.

a) Black-box approach: This approach relies on statistical models where certain variable inputs are selected and based on their interaction/ non-interactions thermal behavior of the buildings is explained. These parameters usually involve weather data, building fabric and system properties. Fair amount of studies has been published by statistical scientists. Majority of these have explained the same with the help of multiple regression model, artificial neural networks, genetic algorithms, etc.

b) Grey-box/parameter estimation models: grey scale models as the name suggests, ascertains key parameters from the physical model to explain the system behavior and further statistical model are developed as a next logical step in determining the end results.

c) Detailed model calibration: the calibration models use detailed law -driven building system simulation modeling results and certain key inputs are tuned manually or automated using machine learning principles to match the measured data. The calibrated models provide in detail explanation of thermal behavior of

buildings and analyze the impact of retrofit packages and prioritize amongst the available set of retrofit packages.

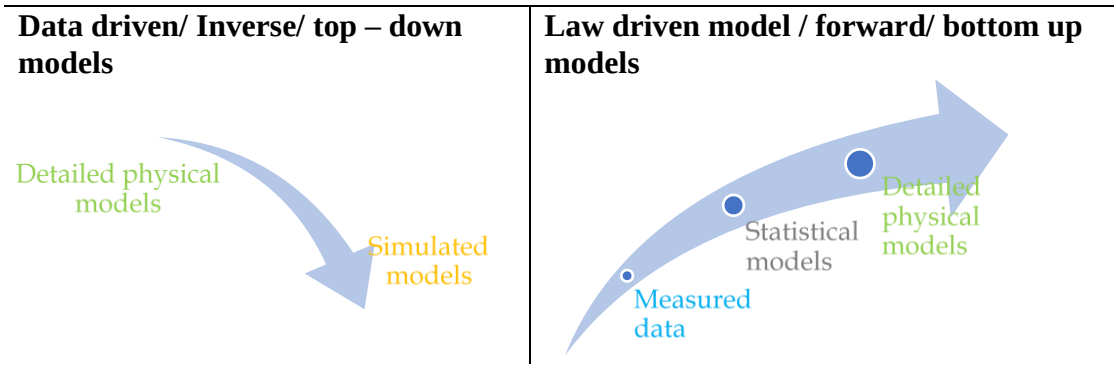

FIGURE 2: TYPE OF BUILDING PERFORMANCE EVALUATION MODELS

Addison (1988) designed and demonstrated a computer aided design methodology suitable for use with any energy simulation program and at any of the development phases. It is a multiple criterion satisficing strategy developed keeping any non-expert building design professional's benefits as central focus for energy efficient buildings and would especially be useful for reaching the critical energy related decisions made early in the programming and conceptual design stages.

Snyder et al (2013) proposed an automated design methodology providing designers a decision support tool rather than an optimization tool, which would generate numerous design alternatives rather than an optimum solution. The study focused mainly on a design of experiments response surface approach and involved very few number of parameters.

Dutta (2013) developed an fully functioning interactive visualization approach termed "Visual Analytics based Decision Support Methodology [VADSM]" which used Multi-Criterion Decision Making (MCDM) regression based models to create dynamic interplays of important variables' alteration affected two performance criteria Energy Use Intensity (EUI) and Peak Energy Demand (PED), while providing a visual range or band of variation of the different design parameters using parallel coordinate representation. It was based on the application of Monte Carlo approaches to create a database of solutions using deterministic whole building energy simulations, along with data mining methods (random forest algorithm) to rank variable importance and reduce the multi-dimensionality of the problem.

Didwania (2015) proposed alternative design methodology to the two prior studies and considering parameter interactions more explicitly and to different types of advanced HVAC systems and their effect in different climates were analyzed and basic VBA based interaction model was created.

The current study dwells into further widening the scope of data analytics in the whole building energy simulations on urban/block level. A way forward would be to reduce the efforts to make explanations based on "bottom-up" prognostic building energy simulation models with the help of previously measured data and find the uncertainty presented by limited number of variable which presents significant change in the behavioral narrative of the target buildings.

2.3 Urban Building Energy Models

2.3.1 Bottom-up Approach

a. Analyzing each building on Individual Basis

Autodesk and ICF international research team (2009) developed a methodology to rapidly estimate energy performance of existing buildings' energy use by using minimal details about the target building. The team focused on digitally capturing the external features of the building and measured data for the internal load profiles. depending upon the confidence interval of data accuracy the model would inherently add 20-30% uncertainty to each parameter.

Joshua et.al. (2012) developed a web-based automated building energy calibration framework called "Autotune" which aimed to replace art with science and expensive human time with cheap computing time. Autotune uses evolutionary computation to calibrate model inputs using any sources of measured data which can map to simulation engine output. An important aspect of the Autotune project is a Trinity Test framework and web service for quantitatively evaluating any calibration algorithm.

b. Dividing the Study Area into Building Archetypes

This kind of approaches are based on building key characteristics like, primary activity, age of building, size of the building, etc. Carlos et.al. (2015) analyzed two deterministic common methods and proposes third probability based method to define uncertain parameters related to building occupancy in the metered data for defining the archetype

for the study area i.e. a residential neighborhood in Kuwait based on the classification of year of construction of the building.

Sokol et.al. (2015) further developed the developed this approach to develop an iterative archetypical model to better ascertain the uncertainties in the metered data models. It was based on Bayesian calibration techniques for the annual and monthly energy usage. The study targeted on accurately modeling end-use differentiation or seasonal variation and argues that aggregated standardization in neither effective nor sufficient to explain the disparity in end-use variations.

Korolija et.al. (2012) developed an archetypal simulation model of office building representing variability a pan-UK office building stock by parameterizing built form, construction elements, occupancy/usage and operational/control strategy. The method is a two-stage process which includes default values suggested for the formulation of the archetype and parametric studies which can be utilized for assessment of energy performance of building stock and evaluating adaptation/retrofitting strategies.

Lara et.al., (2015) adopts cluster analysis algorithms to find out a few school buildings representative of a sample of about 60 schools in the province of Treviso, North-East of Italy, thus reducing the number of buildings to be analyzed in detail to optimize the energy retrofit measures. The study utilized real consumption data of the scholastic year 2011–2012. The data were correlated to buildings characteristics through regression and the parameters with the highest correlation with energy consumption levels used in cluster analysis to group schools. This method supported the definition of representative

architectural types and the identification of a small number of parameters determinant to assess the energy consumption for air heating and hot water production.

Tsanas et.al. (2012) developed a statistical machine learning framework to study the effect of eight input variables i.e. relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area and its distribution on two output variables namely heating load and cooling load for a shoe box model for residential building type. The study systematically investigated the association strength of each input variable with each output variables using classical and non-parametric statistical analysis tools. The research established use of machine learning algorithms for estimating building parameters as a convenient and accurate approach. The study assumed that the actual data bears resemblance with the training dataset of the mathematical model.

3. METHODOLOGY & THEORY

3.1 Introduction

The thesis proposes a new methodology to facilitate the generation & evaluation of building prototypes necessary for reduced computation efforts and effective evaluation of alternatives subject to user-defined target criteria. This methodology is pertinent towards explaining building energy performance at a city or neighborhood scale.

Machine Learning based Prototype Definition Methodology (MLPDM): (FIGURE 3)

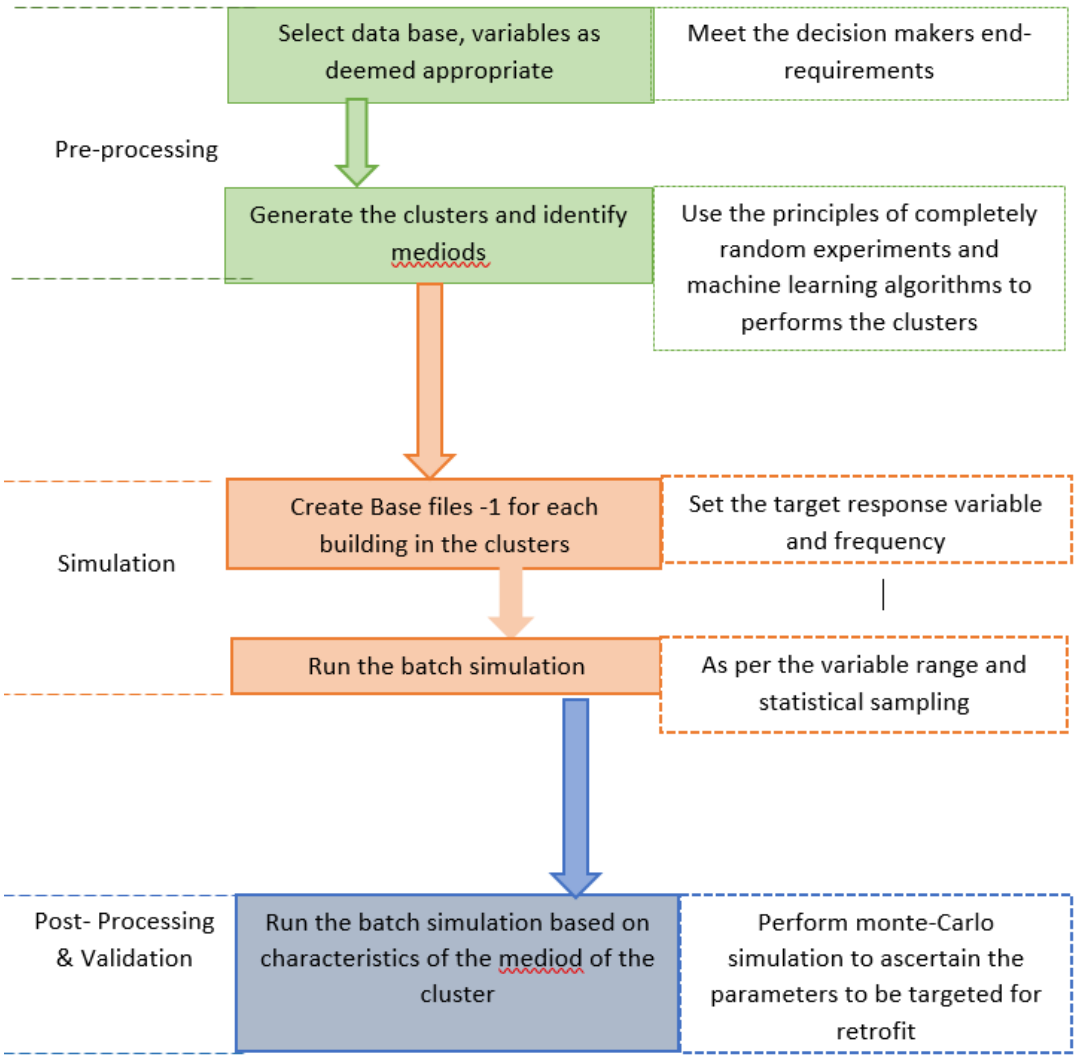

FIGURE 3: MLPDM FLOWCHART: PREPROCESSING- SIMULATION-POST PROCESSING

Stage 1: Pre-processing: Select and identify independent experimental design input variables

The pre-processing of the dataset consists several steps. The experimental design set-up is based on the statistically cleansed datasets, identifying design variables and their logical variability ranges. The cleansing of the dataset is done by removing/ imputing missing data points and outliers. For identifying important design variables, building-type, climatic conditions, the end-goals of the project, statistical algorithms being used

are essential focus areas. Through regression and classification algorithms the relationship between predictor variables and response variables is established. Based on this linear/non-linear relationship variable ranges are identified. The variable importance is established using the supervised or unsupervised learning algorithms such as least squares, random forest, support vector machines, etc. Based on these important variables, appropriate clustering is performed on the dataset which can be K-mean clustering, hierarchical clustering or advanced clustering techniques. The number of factors and their statistical ranges make the possible evaluative combinations range from thousands to millions. To effectively represent the reduced feasible number of representative combinations, appropriate design sampling technique such as Sobol sequences or Latin hypercube or random sampling is necessary. Since, the preprocessing requires user discretion (manual process) towards the project end goals apart from automated experimental design application, this stage may be considered semi-automated.

Stage 2: Simulation – for each cluster – based on their governing characteristics

Selected buildings of each cluster can be now put into a whole building energy simulation program for creating the base file and further conducting the batch simulation, depending on the variable ranges and their uncertainty sampling. The response variable of the target goal can be a direct result of the simulation program output or can be derived amongst the possible output extracted from the program. The direct responses can the zone-level, system-level or facility level energy consumption patterns and derived variables can be the energy usage intensity at site or source and the time-series associated with it. or the source energy usage. This stage has full potential to be fully automated with very few

manual interventions for handling multiple batch-file processing, communicating the input-output variabilities and storing the responses on an online/offline central database.

Stage 3: Post Processing and Validation

The post processing stage involves monte-carlo simulation of the predictor variables and response variables which would provide the bases for selection of appropriate measures towards achieving the set targets. For the validation of the selected process, probability distribution functions are identified for each building and the mean/peak of these Probability Distribution Functions(PDFs). This mean PDF is then compared with the PDF of the mediod representative building.

3.2 Experimental Design

Any design of experiments exercise is focused on understanding the underlying relationship between the predictor variables and response variable. The aim with which the statistical analysis is conducted and what are types of the variables needing to be studied defines the principles of the analysis (FIGURE 4) Any such exercises can be divided into three discrete questions: why and which factors to be studied, what is their individual variability and lastly to what level they are correlated. Dutta et.al. (2013) explained ways to conduct the experiment pertaining to the nonlinear behavior amongst the predictors and response variables and how central composite design techniques are necessary and sufficient to explain the non-linear relationship between building energy

performance indicators. Present study adapts to that approach and tries to explain the urban scale building energy performance criterion.

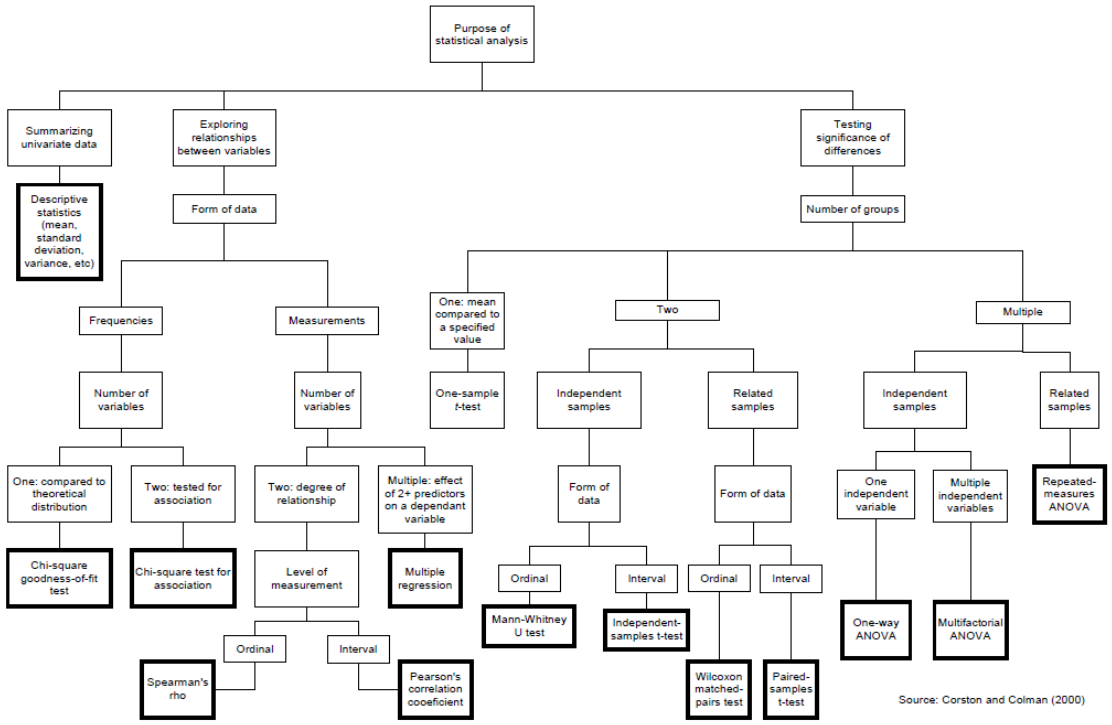

FIGURE 4: PRINCIPLE OF STATISTICAL ANALYSIS – DECISION MAKING TREE FOR CONDUCTING STATISTICAL ANALYSIS OF MULTI-VARIABLE BASED REGRESSION ANALYSIS

3.2.1 Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), Andrienko et.al. (2005) which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and

making transformations of variables as needed. In 1961, Tuckey defined data analysis as: “Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." The main reasons for using EDA are as follows;

- Detection of mistakes
- Checking of assumptions
- Preliminary selection of appropriate models
- Determining relationships among the explanatory variables, and
- Assessing the direction and rough size of relationships between explanatory and outcome variables. In short, EDA gives useful insights into the dataset without including formal statistical modeling.

3.3 Random Forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Brieman). Brieman attempted to improve the conventional bagging algorithm of CART and came up with much stable algorithm. He argued that the instability of CART models’ predictors can be stabilized by making many predictions using multiple weak learners that together constitute an ensemble learner (Breiman, 1998). RF works by building an ensemble of decision trees on bootstrapped samples wherein each tree split is chosen from a limited set of randomly selected features. Since it includes many trees, this ensemble is called a forest. Breiman showed that the accuracy of RF is as good as, or sometimes better than that of SVMs (Breiman, 2001). One of the

reasons why RF is so effective for complex response functions is that it capitalizes on very flexible fitting procedures that can respond to highly local features of the data. Such flexibility is desirable because it can substantially reduce the bias in the fitted values compared to the fitted values from parametric regression. The flexibility in RF comes, in part, from individual trees that can find nonlinear relationships and interactions. Another source of the flexibility is large trees that are not precluded from having very small sample sizes in their terminal nodes. RF consciously address over-fitting by using OOB observations (explained below) to construct the fitted values and measures of fit and by averaging over trees. Yet another source of flexibility is the random sampling of predictors. This strategy allows predictors that work well, but only for a very few observations, the opportunity to participate. This also reduces competition between correlated predictors, and given a large enough number of trees each gets a chance to contribute. This two-part strategy – flexible fitting functions and averaging over OOB observations is highly effective and has the potential to break the bias-variance tradeoff (Berk, 2008).

3.3.1 Random Forest Algorithm

Let $Dn = \{(\boldsymbol{X}i, Yi) : i = (1,23 \cdots N)\}$ where $\boldsymbol{X_i^{(1)}}, \dots, \boldsymbol{X_i^{(d)}} \in \boldsymbol{R^d}, \boldsymbol{Y_i}$ be the independent and identically distributed (i.i.d.) raining data set. Then the Random Forest algorithm suggested by Breiman is constructed as follows (Bae, 2008):

Step 1: Draw K independent bootstrap samples from $B_k, k = 1, \dots, k$, from D_n, where $|B_i| = n$. Note that each consists B_k of n samples chosen randomly from D_n with replacement and |A| is the number of elements in set A.

Step 2: For each $B_k, k = 1, \dots, k$, grow a tree with following rules.

2.1 At each node, randomly select a subset of F variables from d variables, where

F ≤ d is a tuning parameter in the Random Forests algorithm.

2.2 At each node, find the best split (feature variable and split point) among the F variables chosen at 2.1.

2.3 Grow trees to a maximum depth without pruning. That is, grow trees until each terminal node contains no more than 5 training data observations in regression and until each terminal node contains data with same class in classification.

2.4 Let $f(x, D_n, \theta_k)$ be the resulting tree predictor where x is a set of feature variables, θ_k is a randomly chosen variable consisting of subsets of feature variables, split points at each node and B_k . Thus $\theta_j, j = 1,2, \dots, k$ are identical independent distributed random variables

Step 3: Define the final Random Forests predictor $f(x, D_n)$ as

$$f(x, D_n) = \frac{1}{K}\sum_{k=1}^{K} f(x, D_n, \theta_k) \quad \text{(Eqn.1)}$$

3.3.2 Out of Bag Observations and forecasting error

In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the forecasting or test error. When sampling randomly from a set of observations to generate a bootstrap training sample for a single tree an average of 36.8% of the observations are not used for building that individual tree. These observations are

considered “out of the bag” or OOB for that tree. The accuracy of a random forest’s prediction can be estimated from these OOB data as

$$\boldsymbol{OOBMSE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(\boldsymbol{y_i}, \overline{\boldsymbol{\widehat{y}_{iOOB}}}\right)^{\mathbf{2}} \quad \text{(Eqn. 2)},$$

where, $\overline{\boldsymbol{\widehat{y}_{iOOB}}}$ denotes the average prediction for the *i*th observation from all trees for which this observation has been OOB, *n* is the data size.

3.3.3 Predicator Importance with RF

In many statistical learning applications, the goal is not only to achieve high prediction accuracy but also to understand the underlying mechanism, or in other words explore how inputs are related to outputs. Finding relevant variables may be one of the ways to understand this. RF provides two approaches to assess predictor importance.

a) Contribution to Model Fit

One approach to measuring predictor importance is to record the decrease in fitting measure (ex. Gini Index) each time a given variable is used to define a split. The sum of these reductions for a given tree is a measure of importance for the variable, when the tree is built. For RF one can average this measure of importance over the set of trees. However, reductions in the fitting criteria ignore the forecasting skill of a model since the fit measures are computed with the training data and not the test data (OOB Data). If one cannot forecast well it means that the model cannot usefully reproduce the empirical world. Moreover, it can be difficult to translate these contributions to fit statistics into practical terms. (Berk, 2008).

3.3.4 RF Tuning Parameters

Despite the complexity of the RF algorithm and the large number of potential tuning parameters, most of the usual defaults work well in practice. The tuning parameters most likely to require some manipulation are the following:

a) Node Size

Unlike in CART, the number of observations in the terminal nodes of each tree in RF can be very small. Software packages like Matlab and R use the default of 5 for regression and 1 for classification. The goal is to grow trees with as little bias as possible. The high variance of individual trees that would result can be tolerated because of the averaging over a large number of such trees.

b) Number of Trees

The number of trees should be chosen based on the cost of computation. In practice 500 trees are often a good compromise and appear commonly in research. One benefit of a large number of trees is that each predictor will have an ample opportunity to contribute, even if very few are drawn for each split.

c) Number of Predictors Sampled

Most statistical software applications (R, Matlab) by default take the square root of the total number of variables for classification, and one third the total number for regression. Breiman suggested starting with the defaults and then trying a few more or less. In practice, large differences in performance are rarely found and selecting a few predictors each time seems to be adequate, provided the number of trees is in the order of 500 or so.

3.4 Clustering

Clustering techniques are generally classified as partitional clustering and hierarchical clustering, based on the properties of the generated clusters (Everitt et al., 2001; Hansen and Jaumard, 1997; Jain et al., 1999; Jain and Dubes, 1988). Partitional clustering directly divides data points into some pre-specified number of clusters without the hierarchical structure, while hierarchical clustering groups data with a sequence of nested partitions, either from singleton clusters to a cluster including all individuals or vice versa. The former is known as agglomerative hierarchical clustering, and the latter is called divisive hierarchical clustering. Both agglomerative and divisive clustering methods organize data into the hierarchical structure based on the proximity matrix. The results of hierarchical clustering are usually depicted by a binary tree or dendrogram, as depicted in FIGURE 5.

The root node of the dendrogram represents the whole data set, and each leaf node is regarded as a data point. The intermediate nodes thus describe the extent to which the objects are proximal to each other; and the height of the dendrogram usually expresses the distance between each pair of data points or clusters, or a data point and a cluster. The ultimate clustering results can be obtained by cutting the dendrogram at different levels (the dashed line in This representation provides very informative descriptions and a visualization of the potential data clustering structures, especially when real hierarchical relations exist in the data, such as the data from evolutionary research on different species of organisms, or other applications in medicine, biology, and archaeology (Everitt et al., 2001; Theodoridis and Koutroumbas, 2006). Compared with agglomerative methods, divisive methods need to consider $2^{N-1} - 1$ possible two - subset divisions for a cluster

with N data points, which is very computationally intensive. Therefore, agglomerative methods are more widely used. As the current study focuses on agglomerative clustering, here explanation for agglomerative clustering is provided in the following section. The common criticism of classical hierarchical clustering algorithms is high computational complexity, which is at least $O(N^2)$. This high computational burden limits their application in large - scale data sets. In order to address this problem and other disadvantage, some new hierarchical clustering algorithms have been proposed, such as BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) (Zhang et al., 1996) and CURE (Clustering Using Representatives) (Guha et al., 1998). These algorithms are beyond scope of the current study and its applicability should be considered to be explored in the future.

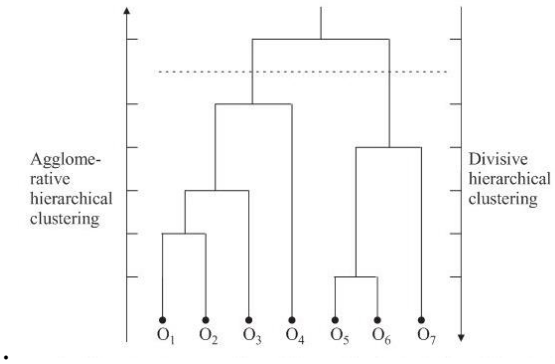

FIGURE 5 AGGLOMERATIVE/DIVISIVE HIERARCHICAL CLUSTER - EXPLANATORY DENDROGRAM – BOTH CLUSTERING IS OPPOSITE TO EACH OTHER AND DATA SET CAN BE DIVIDED INTO PARTS BY CUTTING THE DENDROGRAM AT APPROPRIATE LEVEL.

The root node of the dendrogram represents the whole data set, and each leaf node is regarded as a data point. The intermediate nodes thus describe the extent to which the objects are proximal to each other; and the height of the dendrogram usually expresses the distance between each pair of data points or clusters, or a data point and a cluster. The

ultimate clustering results can be obtained by cutting the dendrogram at different levels (the dashed line in FIGURE 5). This representation provides very informative descriptions and a visualization of the potential data clustering structures, especially when real hierarchical relations exist in the data, such as the data from evolutionary research on different species of organisms, or other applications in medicine, biology, and archaeology (Everitt et al., 2001; Theodoridis and Kou- troumbas, 2006). Compared with agglomerative methods, divisive methods need to consider $2^{N-1}-1$ possible two - subset divisions for a cluster with N data points, which is very computationally intensive. Therefore, agglomerative methods are more widely used. As the current study focuses on agglomerative clustering, here explanation for agglomerative clustering is provided in the following section. The common criticism of classical hierarchical clustering algorithms is high computational complexity, which is at least $O(N^2)$. This high computational burden limits their application in large - scale data sets. In order to address this problem and other disadvantage, some new hierarchical clustering algorithms have been proposed, such as BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) (Zhang et al., 1996) and CURE (Clustering Using Representatives) (Guha et al., 1998). These algorithms are beyond scope of the current study and its applicability should be considered to be explored in the future.

3.4.1 Agglomerative Hierarchical Clustering

General Agglomerative Hierarchical Clustering starts with N clusters, each of which includes exactly one data point. A series of merge operations is then followed that eventually forces all objects into the same group. The general agglomerative clustering

can be summarized by the following procedure, which is also summarized in **Error! Reference source not found.**.

1.Start with N singleton clusters. Calculate the proximity matrix (usually based on the distance function) for the N clusters;

2. In the proximity matrix, search the minimal distance $D(Ci, Cj) = \min_{1 \le m,l \le N, m \ne l} D(Cm, Cl)$, where $D(.,.)$is the distance function discussed later in the section, and combine cluster Ci and Cj to form a new cluster Cij;

3. Update the proximity matrix by computing the distances between the cluster Cij and the other clusters;

4. Repeat steps 2 and 3 until only one cluster remains.

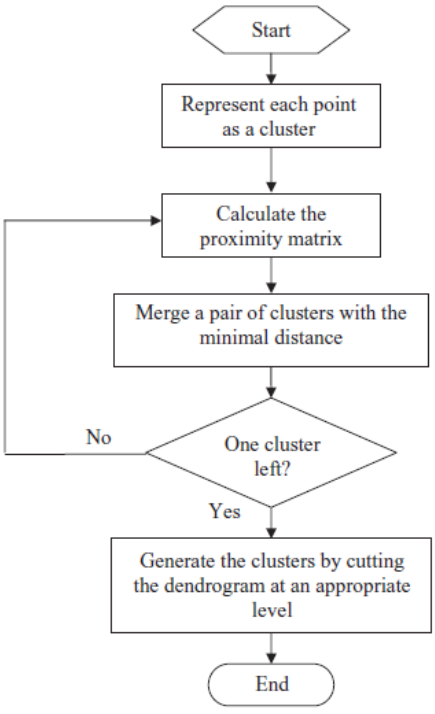

FIGURE 6 AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM FLOWCHART. AGGLOMERATIVE CLUSTERING CONSIDERS EACH DATA POINT AS A CLUSTER IN THE BEGINNING. TWO CLUSTERS ARE THEN MERGED IN EACH STEP UNTIL ALL OBJECTS ARE FORCED INTO THE SAME GROUP.

3.4.2 Clustering Linkage

Several different clustering methods are available. Ward's minimum variance method aims at finding compact, spherical clusters. The complete linkage method finds similar clusters. The single linkage method (which is closely related to the minimal spanning tree) adopts a ‘friends of friends’ clustering strategy. The other methods can be regarded as aiming for clusters with characteristics somewhere between the single and complete link methods. Note however, that methods "median" and "centroid" are not leading to a monotone distance measure, or equivalently the resulting dendrograms can have so called inversions or reversals which are hard to interpret.

Obviously, the merge of a pair of clusters or the formation of a new cluster is dependent on the definition of the distance function between two clusters. There exists a detailed distance definition between a cluster Ci and a new cluster Ci formed by the merge of two clusters Ci and Cj, which can be generalized by the recurrence formula proposed by Lance and Williams (1967) as

$$D\left(C_i,\left(C_i,C_j\right)\right) = \alpha_i D(C_l, C_i) + \alpha_j D(C_l, C_j) + \beta D(C_l, C_j) + \gamma\left|D(C_l, C_i) - D(C_l, C_{ij})\right|, \qquad \textit{Eqn (3)}$$

Where, $D(.,.)$ is the distance function and $\alpha_i, \alpha_j, \beta, \gamma$ are coefficients that take values dependent on the scheme used. The parameter values for the commonly used algorithms are summarized in Table (FIGURE 7), which are also given in Everitt et al. (2001), Jain and Dubes (1988), and Murtagh (1983).

Clustering algorithms	α_i	α_j	β	γ
Single linkage (nearest neighbor)	1/2	1/2	0	−1/2
Complete linkage (farthest neighbor)	1/2	1/2	0	1/2
Group average linkage (UPGMA)	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Weighted average linkage (WPGMA)	1/2	1/2	0	0
Median linkage (WPGMC)	1/2	1/2	−1/4	0
Centroid linkage (UPGMC)	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$\frac{-n_i n_j}{(n_i+n_j)^2}$	0
Ward's method	$\frac{n_i+n_l}{n_i+n_j+n_l}$	$\frac{n_j+n_l}{n_i+n_j+n_l}$	$\frac{-n_l}{n_i+n_j+n_l}$	0

FIGURE 7 LANCE AND WILLIAMS' PARAMETERS FOR AGGLOMERATIVE HIERARCHICAL CLUSTERING. AI, AJ, B, AND Γ ARE PARAMETERS DEFINED IN EQUATION ABOVE. N I, N J, AND N L ARE THE NUMBER OF DATA POINTS IN CLUSTER C I, C J, AND C L, RESPECTIVELY

The single linkage algorithm (Everitt et al., 2001; Johnson, 1967; Jain and Dubes, 1988; Sneath, 1957). For single linkage, the distance between a pair of clusters is determined by the two closest objects to the different clusters. So, single linkage clustering is also called the nearest neighbor method. Following the parameters identified in *Eqn (3)* becomes,

$$D\left(C_i,\left(C_i,C_j\right)\right)=min\left(D\left(C_l,C_i\right),D\left(C_l,C_j\right)\right), \qquad \textit{Eqn (3)}$$

Therefore, the distance between the newly generated cluster and the old one is dependent on the minimal distance of $D(C_l, C_i)$and $D(C_l, C_j)$. Single linkage clustering tends to generate elongated clusters, which causes the chaining effect (Everitt et al., 2001). As a result, two clusters with quite different properties may be connected due to the existence of noise. However, if the clusters are separated far from each other, the single linkage method works well.

The complete linkage algorithm (Everitt et al., 2001; Jain and Dubes, 1988; Sorensen, 1948): In contrast to single linkage clustering, the complete linkage method uses the

farthest distance of a pair of objects to define inter-cluster distance. In this case, *Eqn (3)* becomes

$$D\left(C_i,\left(C_i,C_j\right)\right)=max\left(D\left(\ C_l\ ,C_i\ \right),D\left(\ C_l\ ,C_j\ \right)\right) \qquad \textit{Eqn (4)}$$

It is effective in uncovering small and compact clusters.

- The group average linkage algorithm, also known as the unweighted pair group method average (UPGMA) (Everitt et al., 2001; Jain and Dubes, 1988; Sokal and Michener, 1958). The distance between two clusters is defined as the average of the distance between all pairs of data points, each of which comes from a different group. *Eqn (ii)* is written as

  $$D\left(C_i,\left(C_i,C_j\right)\right)=\frac{1}{2}\left(D\left(\ C_l\ ,C_i\ \right),D\left(\ C_l\ ,C_j\ \right)\right) \qquad \textit{Eqn (5)}$$

The distance between the new cluster and the old one is the average of the distances of $D\left(\ C_l\ ,C_i\ \right)$and $D\left(C_l\ ,C_j\ \right)$.

- The weighted average linkage algorithm is also known as the weighted pair group method average (WPGMA) (Jain and Dubes, 1988; McQuitty, 1966). Similar to UPGMA, the average linkage is also used to calculate the distance between two clusters. The difference is that the distances between the newly formed cluster and the rest are weighted based on the number of data points in each cluster. In this case, *Eqn (3)* is written as

  $$D\left(C_i,\left(C_i,C_j\right)\right)=\frac{n_i}{n_i+n_j}\left(D\left(\ C_l\ ,C_i\ \right)\right)+\frac{n_j}{n_i+n_j}\left(D\left(\ C_l\ ,C_j\ \right)\right) \qquad \textit{Eqn (6)}$$

The centroid linkage algorithm, also known as the unweighted pair group method centroid (UPGMC) (Everitt et al., 2001; Jain and Dubes, 1988; Sokal and Michener, 1958). Two clusters are merged based on the distance of their centroids (means), defined as

$$m_i = \frac{1}{n_i} \sum_{x=C_j} x$$

Where, $n_i$is the number of data points belonging to the cluster. *Eqn (ii)* now is written as

$$D\left(C_i, \left(C_i, C_j\right)\right) = \frac{n_i}{n_i+n_j}\left(D\left(C_l, C_i\right)\right) + \frac{n_j}{n_i+n_j}\left(D\left(C_l, C_j\right)\right) -$$

$$\frac{n_i n_j}{\left(n_i+n_j\right)^2}\left(D\left(C_i, C_j\right)\right), \qquad \textit{Eqn (7)}$$

This definition is equivalent to the calculation of the squared Euclidean distance between the centroids of the two clusters,

$$D\left(C_i, \left(C_i, C_j\right)\right) = \left\|\mathbf{m}_{\boldsymbol{l}} - \mathbf{m}_{(\boldsymbol{ij})}\right\|^2$$

The median linkage algorithm, also known as the weighted pair group method centroid (WPGMC) (Everitt et al., 2001; Gower, 1967; Jain and Dubes, 1988). The median linkage is similar to the centroid linkage, except that equal weight is given to the clusters to be merged. *Eqn (ii)* is written as

$$D\left(C_i, \left(C_i, C_j\right)\right) = \frac{1}{2}\left(D\left(C_l, C_i\right)\right) + \frac{1}{2}\left(D\left(C_l, C_j\right)\right) - \frac{1}{4}\left(D\left(C_i, C_j\right)\right) \qquad \textit{Eqn (8)}$$

This is a special case when the number of data points in the two merging clusters is the same.

Ward' s method, also known as the minimum variance method (Everitt et al., 2001; Jain and Dubes, 1988; Ward, 1963). The object of Ward's method is to minimize the increase of the within - class sum of the squared errors,

$$E = \sum_{K=1}^{K} \sum_{x_i \in C_K} \left\| \boldsymbol{X_i} - \boldsymbol{m_{(k)}} \right\|^2 \qquad \textit{Eqn (9)}$$

where K is the number of clusters and m k is the centroid cluster C k as defined in *Eqn (8)*, caused by the merge of two clusters. This change is only computed on the formed cluster and the two clusters to be merged, and can be represented as

$$\Delta E_{ij} = \frac{n_i n_j}{(n_i + n_j)^2} \left\| \boldsymbol{m_i} - \boldsymbol{m_j} \right\|^2, \qquad \textit{Eqn (10)}$$

Single linkage, complete linkage, and average linkage consider all points of a pair of clusters when calculating them inter - cluster distance, and they are also called graph methods. The others are called geometric methods because they use geometric centers to represent clusters and determine their distances. Everitt et al. (2001) summarize the important features and properties of these methods, together with a brief review of experimental comparative studies. Yager (2000) discusses a family of inter - cluster distance measures, based on the generalized mean operators, with their possible effect on the hierarchical clustering process.

3.5 Building Energy Simulation

For performing the whole building energy simulation

3.5.1 Theory

EnergyPlus is an energy analysis and thermal load simulation program, with its root from BLAST (Building Loads Analysis and System Thermodynamics) and DOE-2 programs.

The principle differences between EnergyPlus and its parent tools during its release were (i) its capability to perform integrated simultaneous simulation where building response is tightly coupled with primary and secondary HVAC systems, (ii) heat balance based solution technique for building thermal loads that allow for simultaneous calculation of radiant and convective effects at both interior and exterior surface during each time step, and (iii) the capability to reduce the time step up to 1minute as against the traditional one hour. There are more advantages of using EnergyPlus, which came at the cost of higher modeling and run times. Over the years, various algorithms have been incorporated within EnergyPlus to allow modeling of novel construction materials like phase change materials, conducting complex shadow analysis due to surroundings, furthermore complex and new HVAC systems, incorporating the details of building automation systems. The calibration of the existing building's energy model involves several steps in addition to the geometry, construction details and HVAC details. It involves measured data for identifying how the building is performing as compared to the design cireteria. This calibration exercise includes parametric analysis of the parameters identified for the retrofit. The steps involved in this process are listed in the FIGURE 8

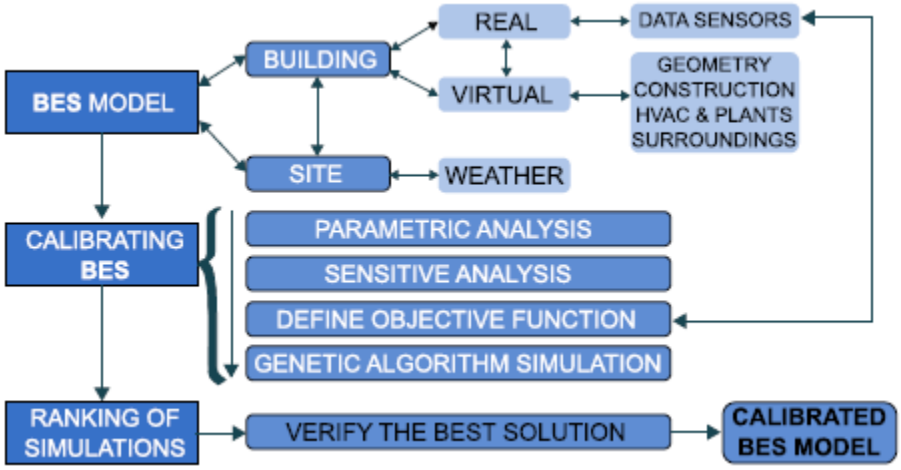

FIGURE 8: SIMULATION STRATEGY - REAL BUILDING TO MODEL CALIBRATION

3.5.2 Sample Building Description

a) Climate selection: The primary goal of the study was to develop retrofit suites for the climate that requires higher cooling energy. Didwania et.al. (2015) conducted similar climate selection study to avoid regions of cold climates that may not have much cooling requirement. Based on the analysis of the study, hot-dry climate of climate zone 3 of the Building America climate Zone 3 (i.e. similar to Phoenix, FIGURE 9) was selected. (IECC-climate guide, 2010).

b) Building Details: For the purpose of creating similar baseline models for each building before introducing the uncertainties based on the cluster details into the energy model, non-geometric details of the ASHRAE 90.1:2010 compliant office building model for Phoenix has been used in the current study. Although, geometric details for each building is different, each building follows following details required for performing the simulation.

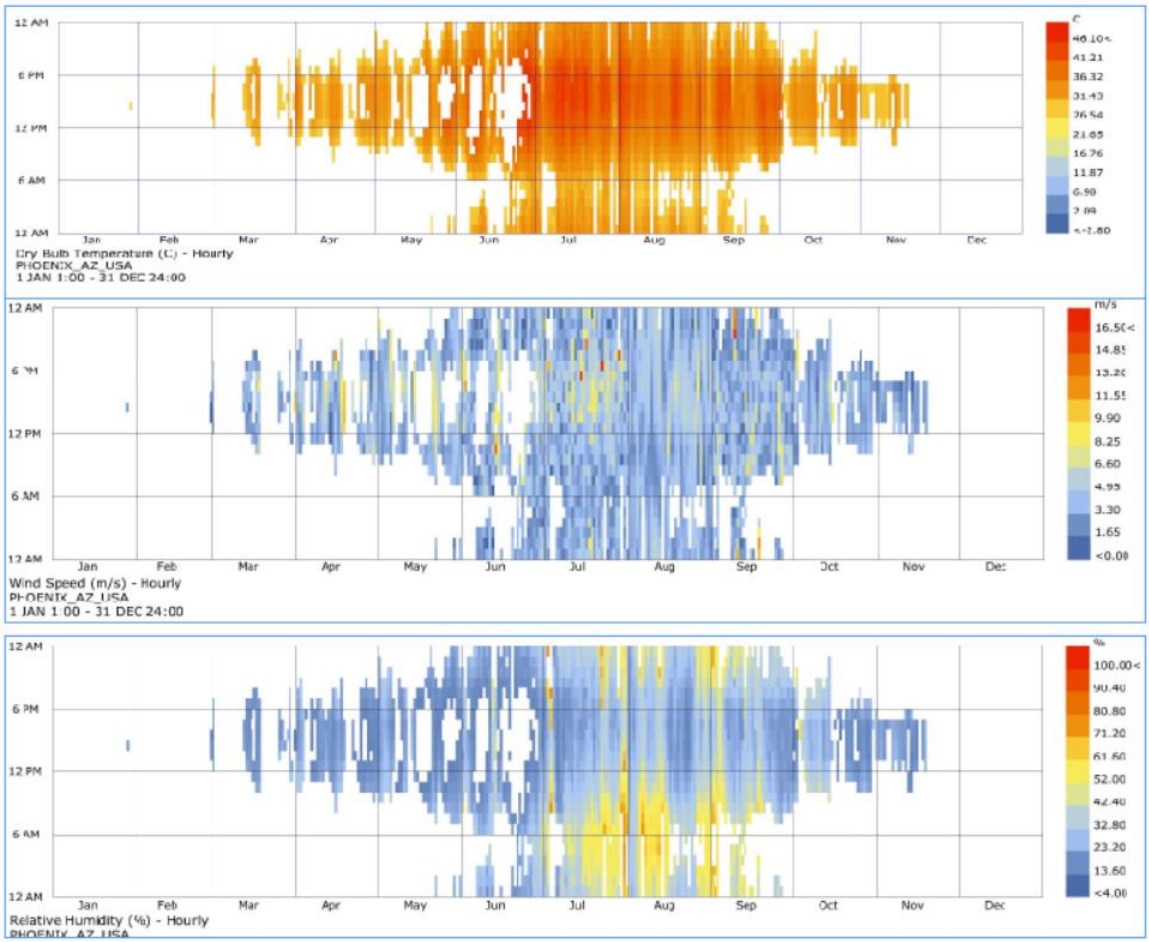

FIGURE 9: TMY2 WEATHER DATA: PHOENIX

PNNL reference Prototype building details: It is a building with rectangular footprint (aspect ratio of 1:1.5) with three floors and total built-up area of 53,600 ft2. The building geometry is shown in FIGURE 10. The base model for each building has 0.33 with the windows distributed uniformly along all four sides of the building. The perimeter zone depth has been modeled as 15 ft., which results in a perimeter area of 40% and a core area of 60%. The zoning is illustrated in FIGURE 11.

The floor-to-floor height is assumed to be 13 ft., with 9 ft. floor-to-ceiling height and 4 ft. plenum. Sill height for the model is assumed to be 3.35 ft.

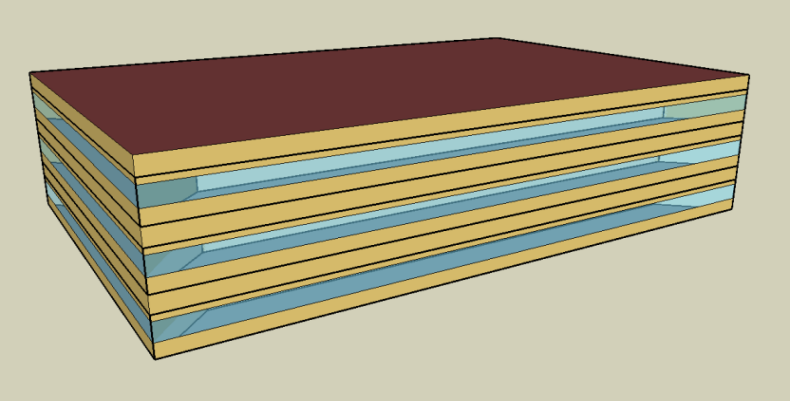

FIGURE 10: PNNL REFERENCE PROTOTYPE OFFICE BUILDING GEOMETRY - COURTESY: PNNL SCORECARD

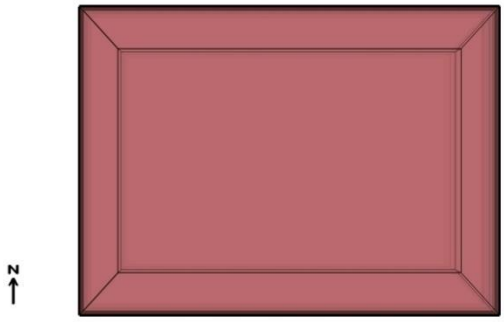

FIGURE 11: ZONING PATTERN OF THE REFERENCE PNNL OFFICE BUILDING PROTOTYPE - COURTESY: PNNL SCORE CARD

All the building characteristics have been modeled to comply with ASHRAE 90.1-2010. For this study, small changes have been made to the models developed by PNNL, in terms of HVAC system the ideal air load system has been assumed to reduce the computation timing of the batch simulations.

c) Building Operational Parameters and Their Ranges

The following section describes the building operation variables considered for the study. The variables are related to building envelope, the internal load i.e. occupancy, lighting, equipment load characteristics. Table 1 assembles the ranges for parameters considered for this study. The entries in bold fonts represent the base case values. Total number of simulation runs for this study is 11,000 since there were 1000 simulations for each

building and 10 such buildings were contained in the sample cluster selected for the analysis. Another 1000 simulation were conducted on the mediod of the building.

TABLE 1: BUILDING PARAMETERS USED FOR THE UNCERTAINTY ANALYSIS

Category	**Building variable Name**	**Definition**	**Range**	**Reference**
Envelope	@@WALLU@@	Wall conductivity	0.2-1.5	(Macdonald, 2002)
	@@WALLABC@@	Wall thermal absorptivity	0.43-0.83	(Macdonald, 2002)
	@@WALLE@@	Wall emissivity	0.87-0.95	(Macdonald, 2002)
	@@ROOFU@@	Roof conductivity	0.2-1.5	(Macdonald, 2002)
	@@ROOFABC@@	Roof thermal absorptivity	0.43-0.83	(Macdonald, 2002)
	@@ROOFE@@	Roof emissivity	0.87-0.95	(Macdonald, 2002)
	@@WINU@@	Glazing material conductivity	1.5-4	(Macdonald, 2002)
	@@WINST@@	Glazing material-solar transmittance	0.16-0.26	(Loutzenhiser, Manz, Moosberger, & Maxwell, 2009)
	@@ACH@@	Air Leakage [air changes per hour]	0.1-1.25	(Heo, 2011)
Internal load	@@EPD@@	Equipment load	0-34	Upper bound based on CBECS cluster
	@@LPD@@	Lighting load	0-17	Upper bound based on CBECS cluster
Controls	@@HSET_OCC@@	Heating set-point	17-25	(Tian & Choudhary, 2011)
	@@CSET_OCC@@	Cooling set-point	17-25	(Tian & Choudhary, 2011)

3.6 Post-Processing and Validation

3.6.1 Monte-Carlo method

Monte Carlo method Statistical method of approximating the solution of complex physical or mathematical systems. The method was adopted and improved by John von

Neumann and Stanislaw Ulam for simulations of the atomic bomb during the Manhattan Project. Because the method is based on random chance, it was named after a gambling resort. [Britannica.com] Monte Carlo methods (or Monte Carlo experiments) are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. Their essential idea is using randomness to solve problems that might be deterministic in principle. They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to use other approaches. Monte Carlo methods are mainly used in three distinct problem classes: [Kroese, et.al.,2014] optimization, numerical integration, and generating draws from a probability distribution. It is a numerical experimentation technique to obtain statistics of output variables of a system computational model, given the statistics of the input variables. In such experiments, the values of the input random variables are sampled based on their distributions, and the output variables are calculated using the computational model. Several experiments are carried out in this manner, and the results are used to compute the statistics of the output variables

a) Theory

Monte Carlo simulation performs sensitivity analysis by building models of possible results by substituting a range of values—a probability distribution—for any factor that has inherent uncertainty. It then calculates results over and over, each time using a different set of random values from the probability functions. Depending upon the number of uncertainties and the ranges specified for them, a Monte Carlo simulation could involve thousands or tens of thousands of recalculations before it is complete. Monte Carlo simulation produces distributions of possible outcome values.

By using probability distributions, variables can have different probabilities of different outcomes occurring. Probability distributions are a much more realistic way of describing uncertainty in variables of a sensitivity analysis. Common probability distributions include:

- Normal – Or "bell curve." The user simply defines the mean or expected value and a standard deviation to describe the variation about the mean. Values in the middle near the mean are most likely to occur. It is symmetric and describes many natural phenomena such as people's heights. Examples of variables described by normal distributions include inflation rates and energy prices.
- Lognormal – Values are positively skewed, not symmetric like a normal distribution. It is used to represent values that don't go below zero but have unlimited positive potential. Examples of variables described by lognormal distributions include real estate property values, stock prices, and oil reserves.
- Uniform – All values have an equal chance of occurring, and the user simply defines the minimum and maximum. Examples of variables that could be uniformly distributed include manufacturing costs or future sales revenues for a new product.
- Triangular – The user defines the minimum, most likely, and maximum values. Values around the most likely are more likely to occur. Variables that could be described by a triangular distribution include past sales history per unit of time and inventory levels.
- PERT- The user defines the minimum, most likely, and maximum values, just like the triangular distribution. Values around the most likely are more likely to

occur. However, values between the most likely and extremes are more likely to occur than the triangular; that is, the extremes are not as emphasized. An example of the use of a PERT distribution is to describe the duration of a task in a project management model.

- Discrete – The user defines specific values that may occur and the likelihood of each. An example might be the results of a lawsuit: 20% chance of positive verdict, 30% change of negative verdict, 40% chance of settlement, and 10% chance of mistrial.

During a Monte Carlo simulation, values are sampled at random from the input probability distributions. Each set of samples is called an iteration, and the resulting outcome from that sample is recorded. Monte Carlo simulation does these hundreds or thousands of times, and the result is a probability distribution of possible outcomes. In this way, Monte Carlo simulation provides a much more comprehensive view of what may happen. It tells you not only what could happen, but how likely it is to happen.

Monte Carlo simulation provides many advantages over deterministic, or "single-point estimate" analysis:

- Probabilistic Results. Results show not only what could happen, but how likely each outcome is.
- Graphical Results. Because of the data a Monte Carlo simulation generates, it's easy to create graphs of different outcomes and their chances of occurrence. This is important for communicating findings to other stakeholders.

- Sensitivity Analysis. With just a few cases, deterministic analysis makes it difficult to see which variables impact the outcome the most. In Monte Carlo simulation, it's easy to see which inputs had the biggest effect on bottom-line results.
- Scenario Analysis: In deterministic models, it's very difficult to model different combinations of values for different inputs to see the effects of truly different scenarios. Using Monte Carlo simulation, analysts can see exactly which inputs had which values together when certain outcomes occurred. This is invaluable for pursuing further analysis.
- Correlation of Inputs. In Monte Carlo simulation, it's possible to model interdependent relationships between input variables. It's important for accuracy to represent how when some factors go up, others go up or down accordingly.

An enhancement to Monte Carlo simulation is the use of Latin Hypercube sampling, which samples more accurately from the entire range of distribution functions.

3.6.2 Latin Hypercube Sampling

Since, the building energy performance can be termed as a multivariate and time dependent, large number of combinations of the variations amongst the predictors need to be effectively considered while performing the uncertainty analysis. Latin Hypercube Sampling provides stratified sampling scheme to improve the k-dimensional input space for large scale simulations such as the urban building energy modeling. Iman (2008) explains how single sample can provide useful information when some input variable (here the building operational variables) dominate certain key responses or time intervals

(thermal behavior of the building). LHS considers the effect of entire range of samples and thus more effective than the random sampling.

a) How it works: Key to LHS is stratification of the probability distribution of

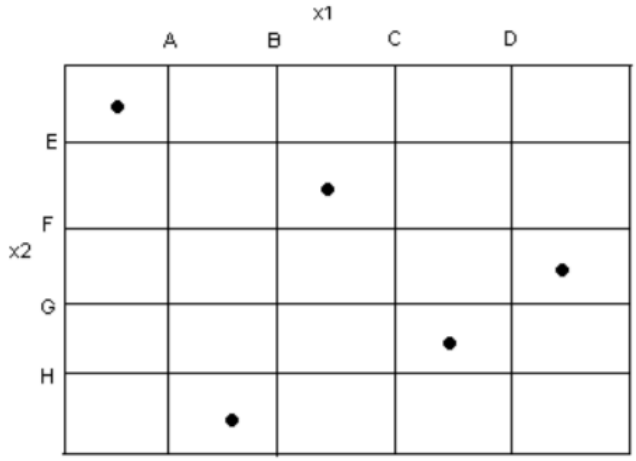

FIGURE 12 LATIN HYPER CUBE SAMPLING- TWO DIMENSIONAL

parameter variability. By stratification, the distribution is divided into equal intervals. From each interval, a sample is taken randomly. **Error! Reference source not found.** explains the process in a simple two-dimensional format. i.e. a square containing a sample position is a Latin square if and only if there is only one sample in each row and column. This can be generalized to any number of dimensions thus Latin hypercube.

4. ANALYSIS AND DISCUSSION

4.1 CBECS Database

The Commercial Buildings Energy Consumption Survey (CBECS) is nationally conducted quadrennially by the U.S. Energy Information Administration to collect basic statistical information about energy consumption and expenditures in U.S. commercial buildings and information about energy-related characteristics of these buildings. The survey is based upon a sample of commercial buildings selected according to the sample design requirements described below. A ‘building,’ as opposed to an ‘establishment,’ is the basic unit of analysis for the CBECS because the building is the energy-consuming unit. Commercial buildings include all the buildings in which at least half of the floor space is used for the purpose that is non-residential, non-industrial or non-agriculture. The CBECS is conducted in two phases, (i) Building Survey and (ii) Energy supplier survey. The most recent survey, CBECS 2012, was the tenth survey conducted since 1979 and is used in this study (EIA, 2017).

The 2012 CBECS target population consisted all commercial buildings that were larger than 1,000 square feet in the U.S. (except for commercial buildings located on manufacturing sites). Unfortunately, the finest level of geographic detail that is publicly available in CBECS is the Census division and Building America Climate region. In addition, building characteristics that could potentially identify a responding building, such as orientation, façade details, etc. are also masked to protect the respondent's identity.

4.2 Data Preprocessing

4.2.1 Data Filtration

The first step of the regression analysis is to filter out observations from the original CBECS database that are of no consequences for the further analyses. Three types of filters are applied sequentially:

1. Building Type Filters: As mentioned above, building use has a significant impact on building energy consumption. Thus, each building type deserves a unique regression model. Current study is limited to office buildings only.

2. Feasibility Filters: Based on prior studies involving similar regression analysis, certain variables have been found to have significant impacts and should be included in the variable selection list. These variables of data samples should indicate 'typical' buildings. For instance, a typical building shall be operated for more than 10 months of a year (PBA = 2); the building shall be air conditioned (percent cooled > 0, percent heated > 0).

3. Outlier Filters: Outlier points shall be eliminated to achieve higher accuracy for common buildings. The criteria used to identify the outliers is to identify the all the buildings whose EUI value lie under the first and third quartile of the total EUI range across the dataset.

The study was carried out by applying these three sets of filters to the original CBECS 2003 micro data and ultimately include 1054 office buildings for this regression analysis Listed in Table 2). In this, the number of sample buildings are those buildings which CBECS selects as unique buildings which is representative of multiple number of buildings. These number is identified by the weightage factor associated with each

unique building. By adding these weightage factors linked with the unique building sample, the total number of representative buildings are calculated. Similarly, the total floor area is calculated. These samples include the buildings surveyed over different climate areas across the U.S.

TABLE 2: SUMMARY OF DATA FILTERS CONSIDERED FOR THE SCREENING CRITERIA

Condition for Including an Observation	Rationale	Number of sample buildings included	Total number of representative buildings	Total floor area of the representative buildings
All data sets	Data source	6720	5,557,138.45	87,067,520,902
PBA =2 i.e. office buildings	Office buildings	1331	983,514	15,596,609,110
Months in use last year >= 10	A typical building being used	1322	971,868	771,618,744.3
Percent cooled > 0, Percent heated > 0	Building must be conditioned	1268	919,884	1,230,250,184
Must have at least 1-person computer	Must be a functional office building	1268	919,884	15,226,490,774
EUI_Primary <= 170 kbtu/sqft/yr	Eliminate outliers outside [Q1-1.5IQR, Q3+1.5IQR]	1191	871,544	14,219,678,539
Floor Area <= 36750 sqft	Eliminate outliers outside [Q1-1.5IQR, Q3+1.5IQR]	1054	867,739	11,437,561,734

As evident from the table 2, the remaining 1054 buildings represent 867,739 actual buildings according to the weighting factors applied to data samples. Considering weighting factors, histograms of gross floor area, climate characteristics, and primary EUI are plotted in the following figures (FIGURE 13,FIGURE 14,FIGURE 15) as part of the univariate exploratory data analysis.

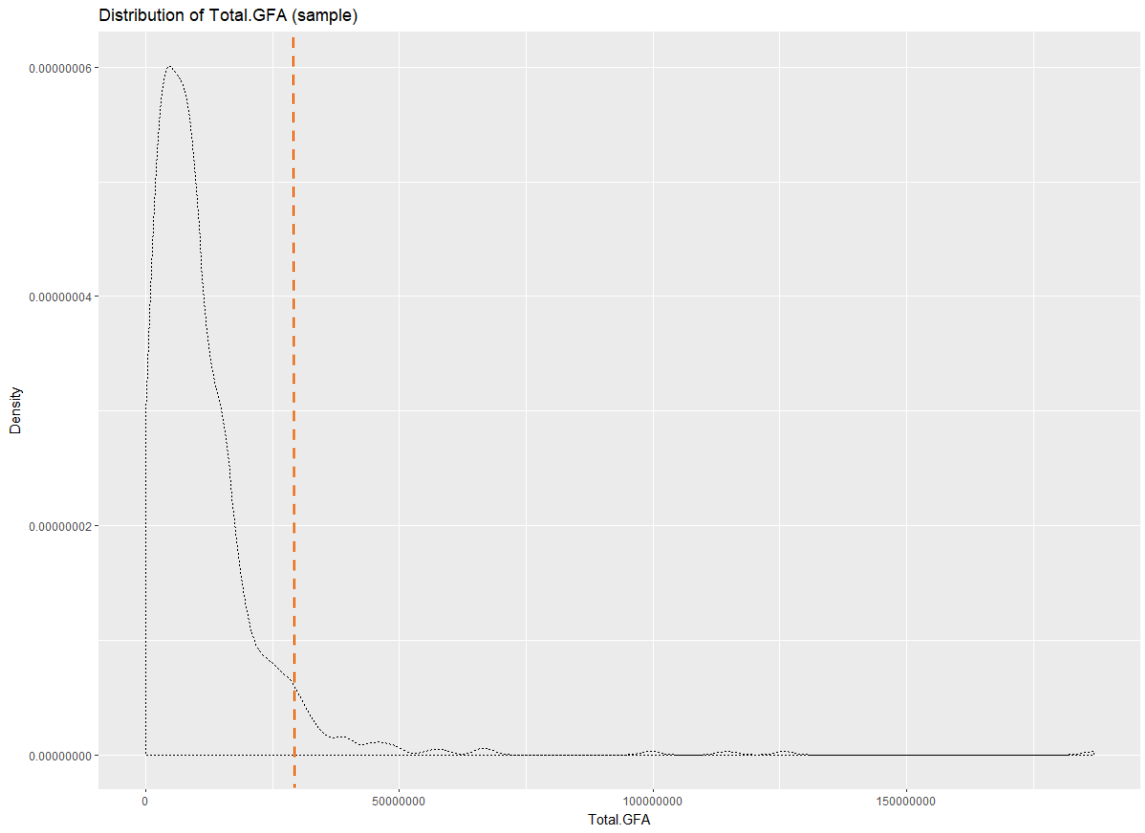

FIGURE 13: GROSS AREA DISTRIBUTION

First step towards studying the EUI distribution patterns is to identify the gross floor area distribution ranges. As evident from the FIGURE 13, over 95 % buildings (marked by the dashed line) are within the range of medium size office buildings defined by the DOE (Reference buildings). Further, as shown in FIGURE 14, the climate variation across this select CBECS dataset is studied based on the value of cooling degree days and heating degree days associated with these buildings. Finally, the EUI distribution is studied for getting initial understanding of how the distribution is. As shown in FIGURE 15, the EUI distribution is skewed left from the standard deviation of the EUI value range across the dataset.

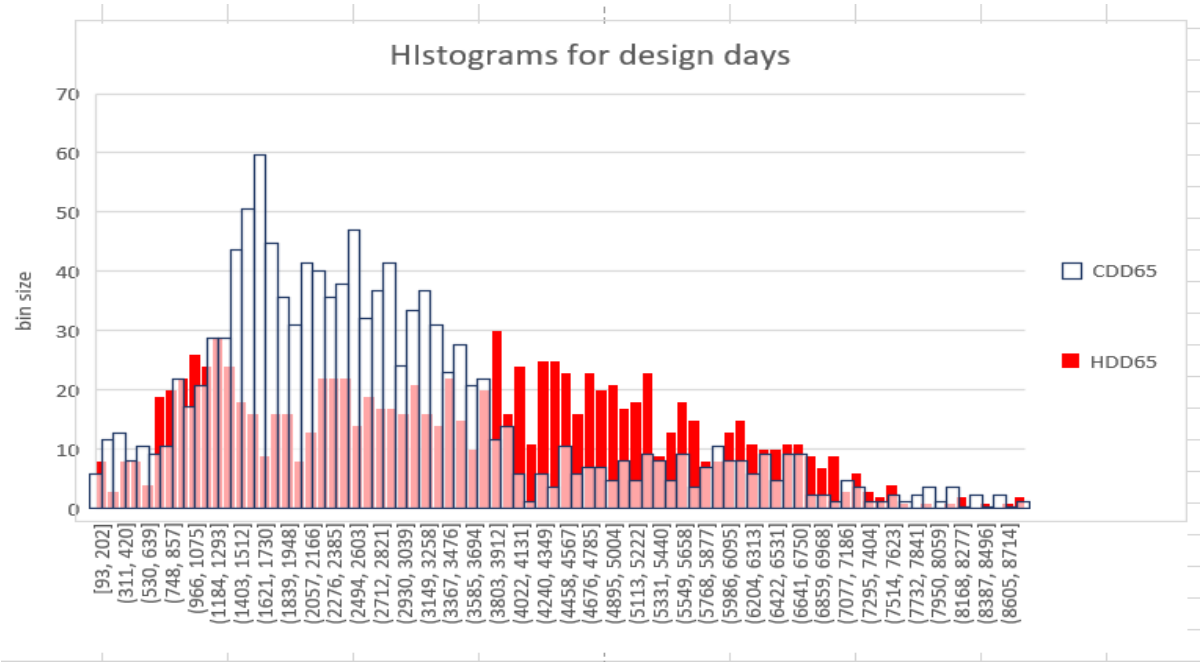

FIGURE 14: HISTOGRAM FOR CDD/HDD

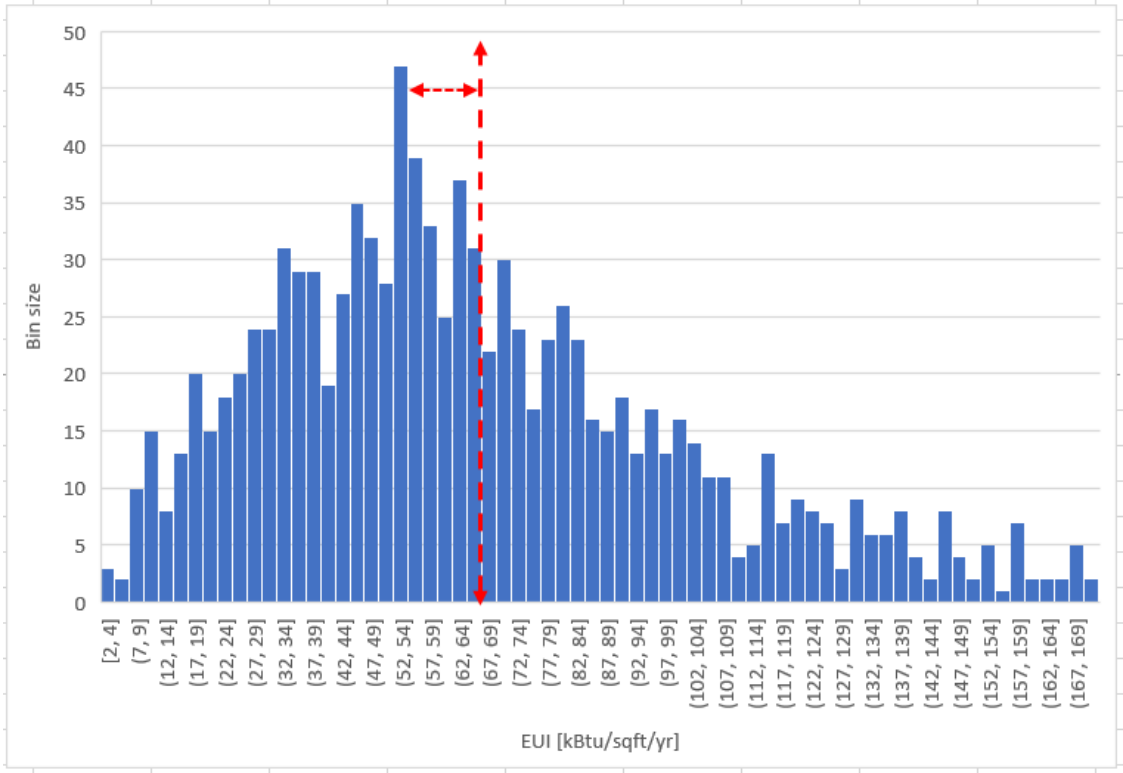

FIGURE 15: PRIMARY EUI DISTRIBUTION

4.2.2 Selection of Variables and Multi Linear Regression Analysis

Many observational variables in CBECS 2012 are potentially relevant to building energy consumption. In this study, 30 variables potentially may have direct impact to the primary EUI and thus are considered as candidates for the variable selection. These parameters are listed in Table 3.

TABLE 3: CBECS DATABASE – VARIABLE SELECTED FOR REGRESSION

Category	CBECS 2003 Variable Name	Definition
Climate	HDD65	Cooling degree days based on 65°F
	CDD65	Heating degree days based on 65°F
	REGION	Census region
	CENCIV	Census division
	PUBCLIM	Building America climate region
Construction	SQFT	Square Footage
	YRCON	year of construction
	DAYLTP	daylight percentage lit
	NFLOOR	number of floors
	FLCEILHT	floor to floor height
	WINTYP	type of window installed
Usage	COOLP	percentage building being cooled
	HEATP	percentage building being heated
	WKHRS	number of hours occupied per week
	MONUSE	number of months in active use per year
	NWKER	number of worker in the building
	PCTERMN	number of computers
	SERVERN	number of servers
	PRNTRN	number of printers
	COPIERN	number of copiers
	RFGWIN	number of refrigerated
	RFGRSN	Number of residential refrigerators
	RFGVNN	number of refrigerated vending machines
	FLUORP	percentage lit by Fluorescent bulb
	CFLRP	percentage lit by CFL
	BULBP	percentage lit by BULB
	HALOP	percentage lit by halogen bulb
	HIDP	percentage lit by high intensity bulb
	LEDP	percentage lit by LED
	OTLTP	percentage lit by other lights

4.2.3 Results of OLS Model of CBECS Data

The objective of this section is to evaluate whether the relationship between EUI and various input parameters. The reason behind conducting this is to critique the behavior explained by linear models in various past studies. These input parameters are related to details of climate region, building envelope, internal loads, controls and schedules.

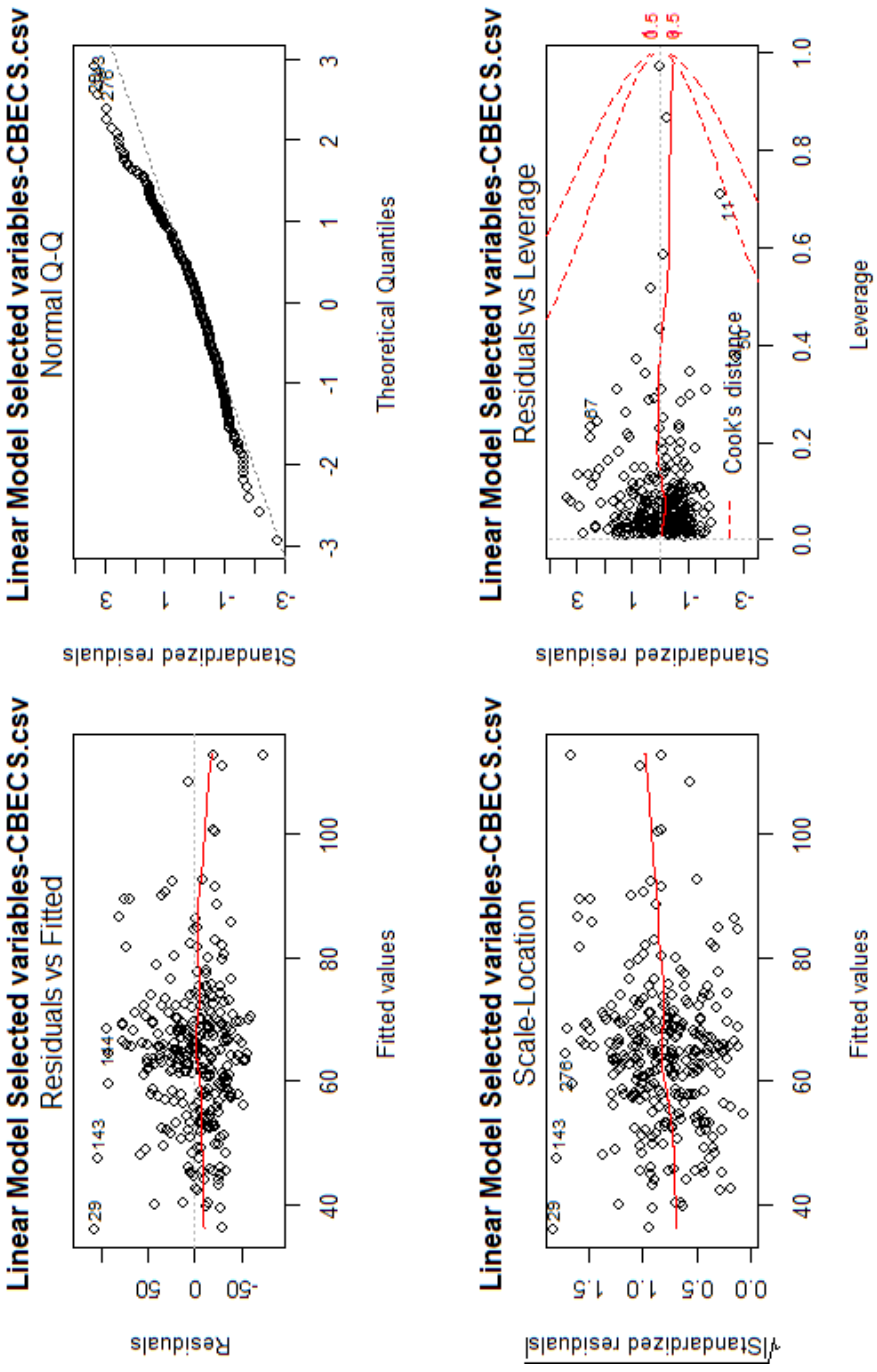

FIGURE 16: OLS MODEL RESULTS

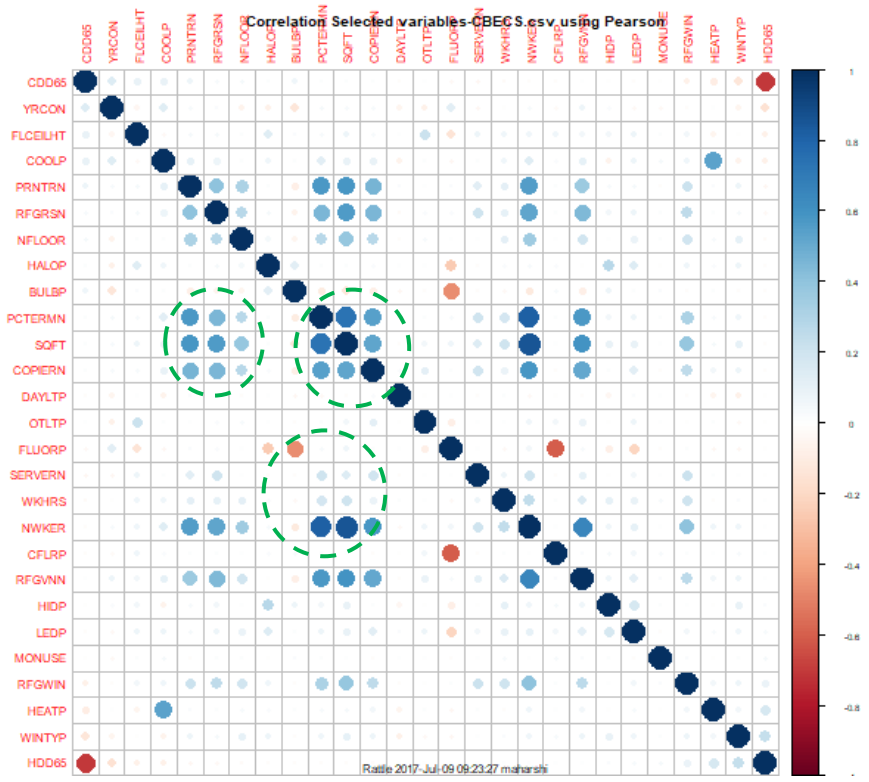

FIGURE 17: CORRELATION MATRIX BASED ON PEARSON COEFFICIENT

As explained in the theory, an OLS model assumes 4 key behaviors:

1. Independence: The OLS assumes that no two input variables are correlated. The coefficient matrix in FIGURE 17 shows the contrary. Certain key variables are highly correlated; such as the encircled region in the figure. The scale of the graph gradient of red to blue with increasing values of the correlation matrix. The graph confirms the generally conceivable understandings regarding few correlations. The higher degree of correlation between the 3 input variables namely, number of floors and equipment, occupants and amount of lighting present is one such example.
2. Linearity: To check whether the linear relation between input variables and response variable, the residuals vs fitted graph is used. the residuals, i.e. vertical distance from the point to the regression line vs fitted values (predicted EUI for each building – y

hat). If no scatter, all fall exactly on the dashed grey line. Ideally, in this case, the red line (smoothed curve) that passes through the actual residuals should be relatively flat, but it is not observed in this case.

3. Normality: One of the assumptions of a least-squares regression is that the errors are normally distributed. QQ plot evaluates this assumption. As it is evident from the figure (upper right graph in FIGURE 16) scatter plot is not so close to the dashed line on both tails, i.e. heavily tailed dataset. To get rid of this anomaly, the data set needs to be transformed and in turn the linear relationship does not remain first order linear regression.
4. Homoscedasticity: As per the definition of homoscedasticity, the variance of EUI should be the same for the given 30 input variables. In general, mild departures do not have significant adverse effects. The residual vs leverage graph can be useful to evaluate this assumption.

a. Leverage: An observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviate from its mean. High leverage points can have a great amount of effect on the estimate of regression coefficients. This is plot on the X axis of the above-mentioned graph (FIGURE 16).
b. Outlier: In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent variable value is unusual given its value on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem. Here, the high leverage points are identified as outliers and should be eliminated from further analysis.

c. Cook's distance (or Cook's D): A measure that combines the information of leverage and residual of the observation. The red lines explain the same. The building data number 67, 11 and 50 are higher leverage or higher residual points in this study and should be further analyzed to understand the reasons behind it.

The variance of the EUI from the graph is evidently not the same and funneled shape. Thus, EUI and its input variables do not show homoscedasticity, and the funnel-like shape shows that it is rather heteroscedasticity.

Based on these outcomes, the dataset is further analyzed to understand the individual relationship of each variable and the target response i.e. EUI. The detailed discussion of the same is provided in the

APPENDIX A. The main findings of this is that individual relationship of each input variable with the EUI (response variable) varies with the variation in climate region as well as for a climate also. For understanding this behavior, the study is further carried out using statistical methods pertaining to non-linear regression modeling. A widely used non-linear regression data mining approach is evaluated and explained in detail in following sections. Before exploring that let's restrict the selection of the buildings which are representative of a particular climate (here, climate 2B -Phoenix)

4.2.4 Selection of building with the climate conditions like Phoenix

After the establishing the basic understanding of the variabilities of the input variables vis-à-vis response variables, the next step is to identify the buildings with the climate

similar to that of the Phoenix city for creating a representative database of an urban neighborhood. The prime reason to select the Phoenix climate area is to focus the study area with dominant cooling requirements and design retrofit policies for this climate. From the database of 1054 unique office buildings 82 buildings are identified for climate similar to Phoenix. These buildings are selected on the basis of degree design days. Buildings with less than 1000 HDD 65 and more than 3000 CDD are selected for the analysis. The variables considered for the random forest are updated from the current set of variables, to incorporate the findings of the preceding data analysis.

4.2.5 Random Forest

The random forest algorithm is utilized to identify the variables amongst the database based on which the clustering is performed. As explained in the theory there are 2 key parameters based on which the stability of the model is dependent. These are called tuning parameters, namely the number of trees and node size. The current database is put in a data frame of R and after setting the appropriate datatypes random forest algorithm (RF algorithm) is performed on the data frame. A snippet of r code is shown in FIGURE 18. The inputs based on which the RF algorithm is tuned are number of tree in each iteration (i.e. “ntree”) and number of branches at each node (i.e. “mtry”). The next step is to optimize the algorithm to suite the requirements of the study i.e. find optimal value for these two above mentioned variables.

```
library(caret)
library(randomForest)
library(ROCR)
set.seed(4)
rf <- randomForest(EUI_total~., data=cbecs.offices,mtry=3, ntree= 500,importance = TRUE, na.action = na.omit)
print(rf)
plot(rf)
```

FIGURE 18 RANDOM FOREST ALGORITHM R CODE

The initial regression model obtained by this code is studied to find how the model is performing i.e. results of the prediction made by the model and actual dataset is studied to make sure the key outcomes of this model used in further analysis are accurate.

The results of the model are shown here in FIGURE 19,

```
Call:
 randomForest(formula = EUI_total ~ .,
              data = crs$dataset[crs$sample, c(crs$input, crs$target)],
              ntree = 500, mtry = 3, importance = TRUE, replace = FALSE, na.action = randomForest::na.roughfix)

               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 791.5365
                    % Var explained: 16.14
```

FIGURE 19 VARIABILITY EXPLAINED BY THE RF ALGORITHM

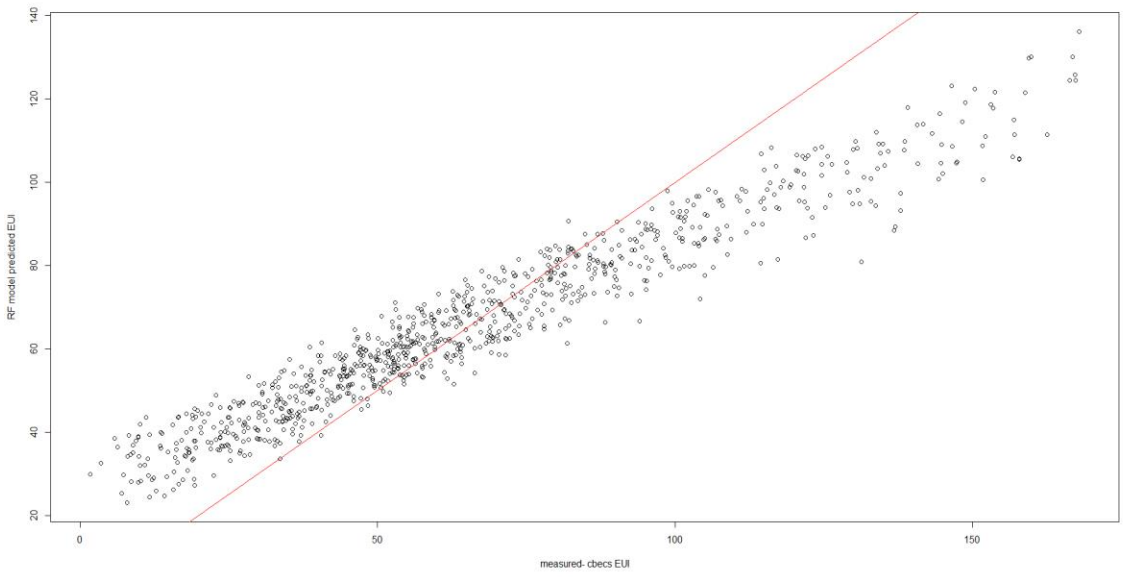

FIGURE 20 CBECS DATA VS RF MODEL PREDICTION

The FIGURE 20 and FIGURE 21 shows the snippets of the R-language code for the random forest algorithm and its results. The results show that model was able to correctly predict only 16%. The mean squared residuals are 800. When considering variable importance study, it is needed to understand that these two outputs are not correct

explanation towards the question. The model prediction values and the actual data set should be compared to study the nature effectiveness of the model prediction. The FIGURE 20 shows the model prediction vs actual data graphically. The model fits the actual dataset effectively to take the variable importance into consideration. The FIGURE 21 explains the number of tree and the error graph, as shown, the error is stable after 200 trees. Thus, one parameter for tuning the RF model is identified based on this, another variable i.e. number of splits at each node should be identified and results of the same are shown in FIGURE 22.

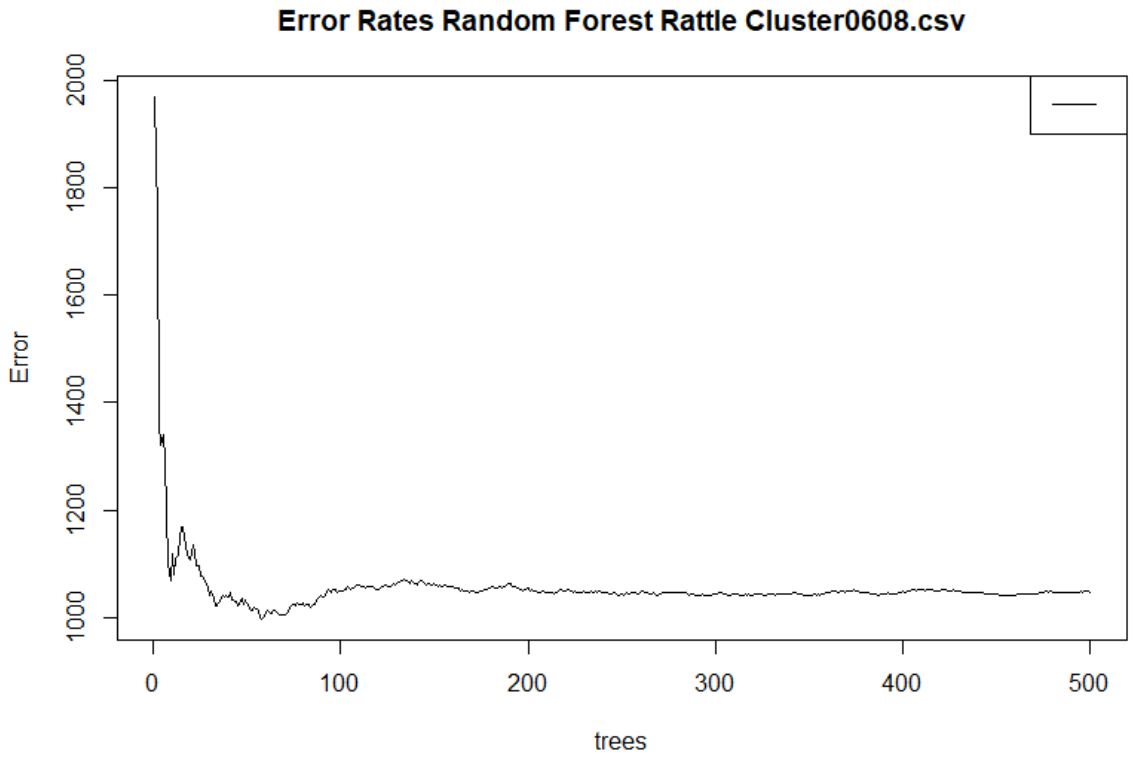

FIGURE 21 : ERROR VS NUMBER OF TREES- RF

The FIGURE 20 and FIGURE 21 shows the snippets of the R-language code for the random forest algorithm and its results. The results show that model could correctly predict only 16%. The mean squared residuals are 800. When considering variable

importance study, it is needed to understand that these two outputs are not correct explanation towards the question. The model prediction values and the actual data set should be compared to study the nature effectiveness of the model prediction. The FIGURE 20 shows the model prediction vs actual data graphically. The model fits the actual dataset effectively to take the variable importance into consideration. The FIGURE 21 explains the number of tree and the error graph, as shown, the error is stable after 200 trees. Thus, one parameter for tuning the RF model is identified based on this, another variable i.e. number of splits at each node should be identified and results of the same are shown in FIGURE 22

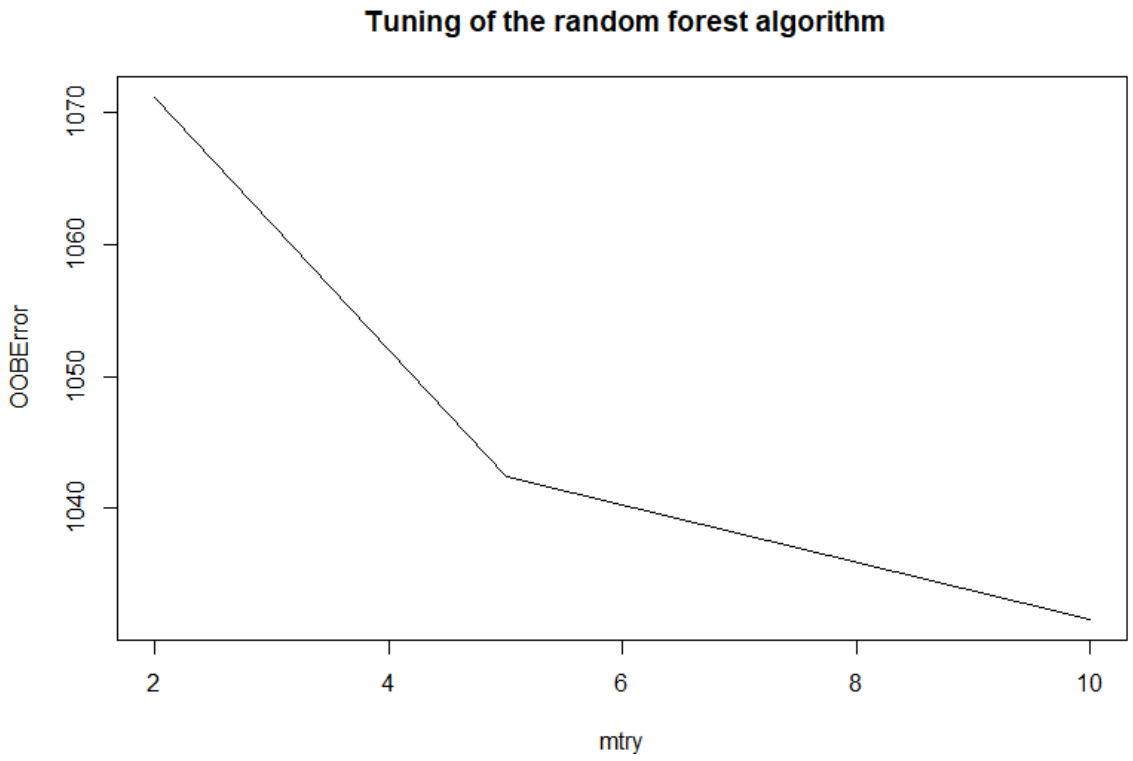

FIGURE 22 TUNING THE RANDOM FOREST ALGORITHM – OPTIMIZE THE VALUE OF MTRY

Based on these optimized RF model, the random forest algorithm is further used for identifying the variable importance for the non-regression. The attribute for identifying the variable importance are explained in detailed in the theory and based on the MSE

attribute the variable importance plot is drawn.

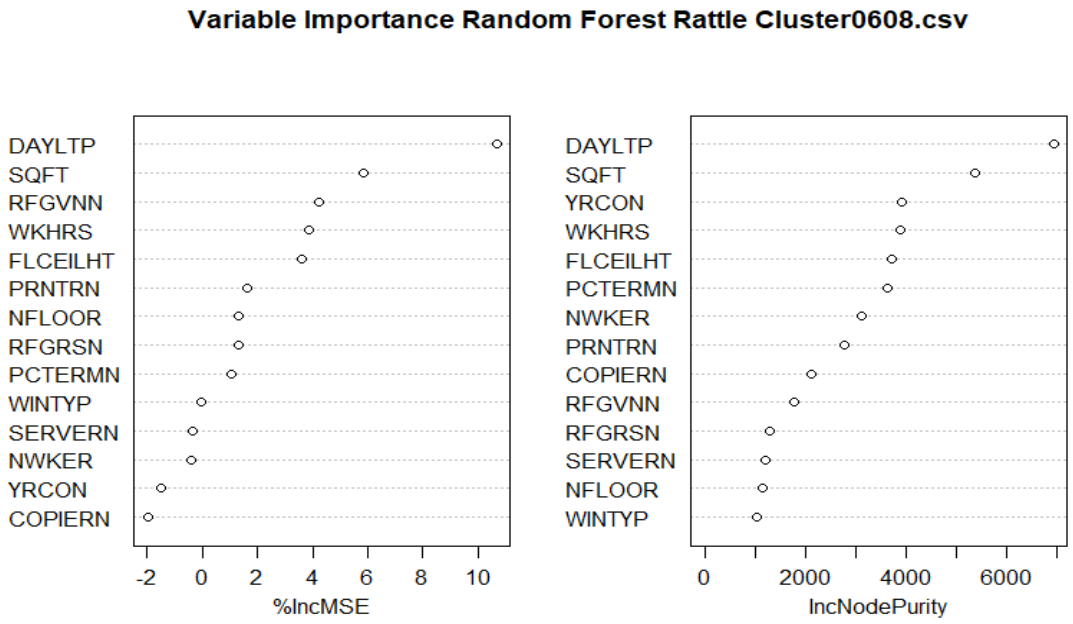

FIGURE 23 RF- VARIABLE IMPORTANCE PLOT

The important variables from the graph are shown on graph FIGURE 23. All the variables with more than |1| % of MSE are of importance and clustering should be carried out based on this parameter. Here, %IncMSE is the most robust and informative measure of any RF model. It is the increase in MSE (mean squared error) of predictions (estimated with out-of-bag-CV) as a result of any variable being permuted(values randomly shuffled).The higher the value of %IncMSE, higher would be the importance of the measure. The graph showing the parameters importance values for each variable on the right is based on the loss function by which best splits are chosen for building the random forest. The loss function is mse for regression and gini-impurity for classification. More useful variables achieve higher increases in node purities, that is to find a split which has a high inter node 'variance' and a small intra node 'variance'. IncNodePurity is biased and should only be used if the extra computation time of calculating %IncMSE is

unacceptable. Since it only takes ~5-25% extra time to calculate %IncMSE, this would almost never happen. Now, based on the select important variables the buildings in the dataset are clustered and studied the potential impact of use of a archetype to represent each cluster of the building.

4.2.6 Hierarchical Clustering

A hierarchical cluster analysis is performed using a set of dissimilarities for the n objects being clustered. Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each stage distances between clusters are recomputed by the Lance–Williams dissimilarity update formula according to the clustering method being used.

As the k means clustering requires the user to specify the number of clusters, and finding the optimal number of clusters can often be hard. Hierarchical clustering is an alternative approach which builds a hierarchy from the bottom-up, and doesn't require us to specify the number of clusters beforehand.

The algorithm works as follows:

- Put each data point in its own cluster.
- Identify the closest two clusters and combine them into one cluster.
- Repeat the above step till all the data points are in a single cluster.
- Once this is done, it is usually represented by a dendrogram like structure.

Once this is done, it is usually represented by a dendrogram like structure.

Present study utilizes the centroid linkage clustering method to determine relative distance between two clusters. This method finds distance between the centroid of each cluster.

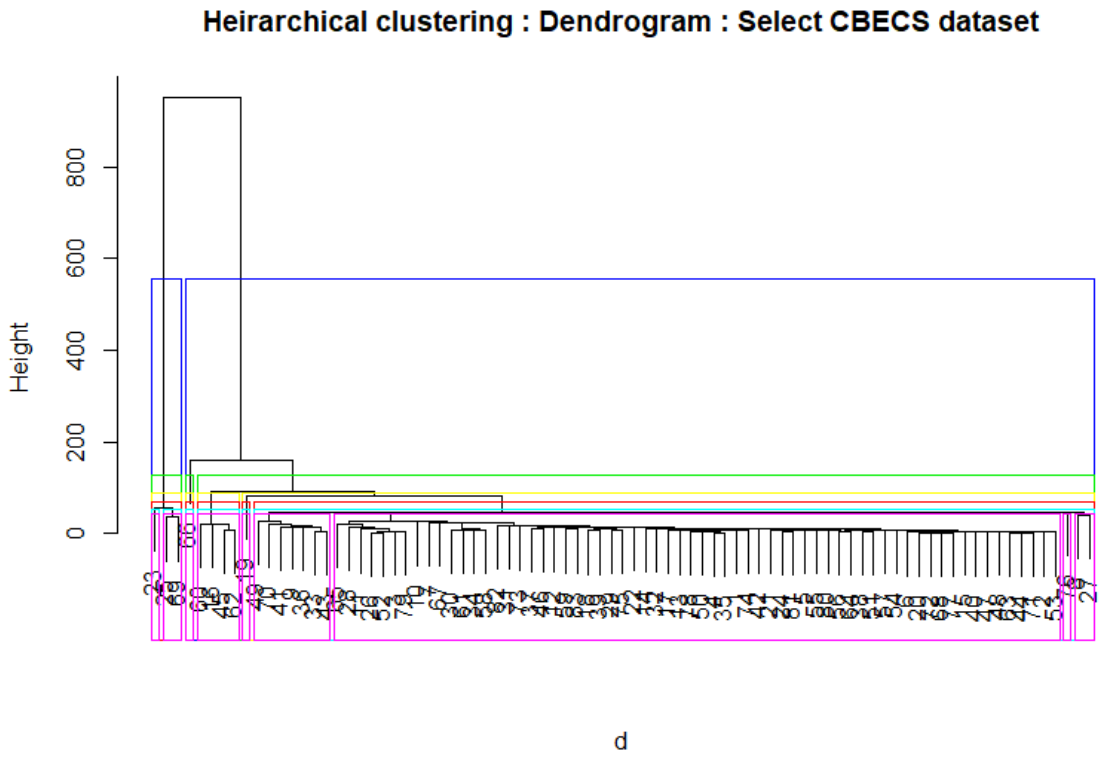

FIGURE 24: HIERARCHICAL CLUSTERING: DENDROGRAM- SELECT CBECS DATASET

The hierarchical clusters' key insights can be drawn from the Dendrogram graph shown (FIGURE 24) above. The present study utilized the Euclidean distance between each cluster's centroid. The dendrogram shows this distance on Y-axis and based on this distance, further the dendrogram can be cut in several branches, terming each branch as one cluster. The above shown dendrogram shows how the dendrogram can be divided

from 2 to 9 clusters. The number of effective clusters were identified based on the change in Euclidean distance between the clusters.

After 9 clusters, up till 30 clusters this Euclidean distance does not change further. So, the study is further continued restricting the number of clusters to 9 only.

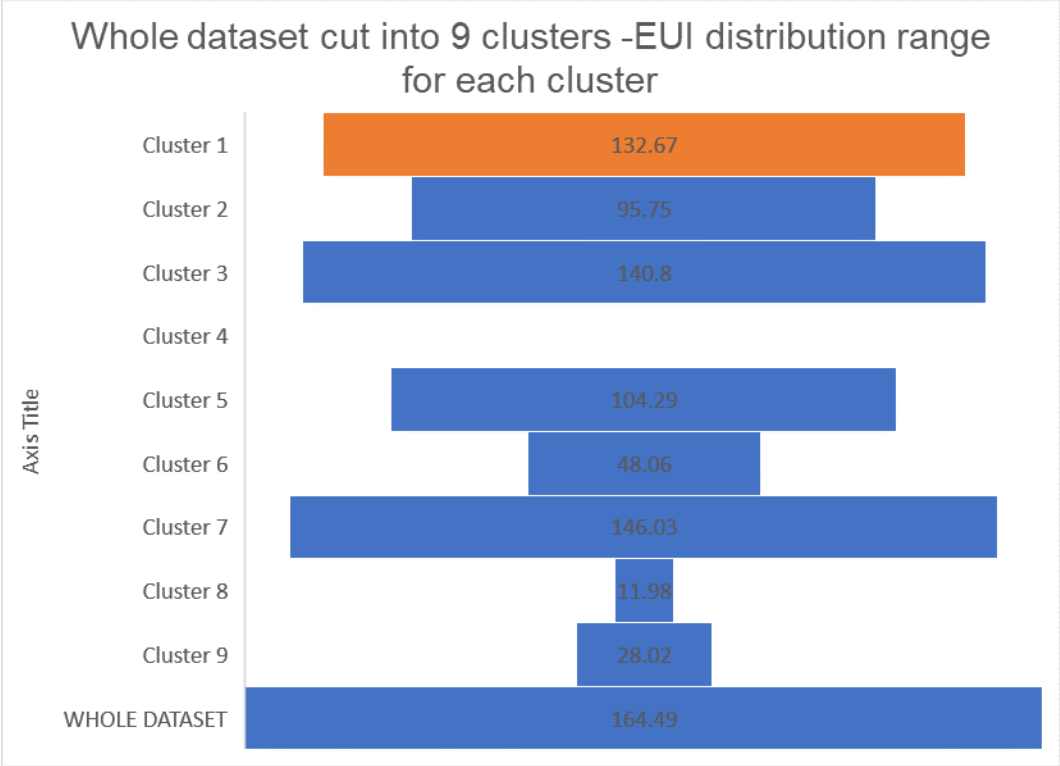

FIGURE 25 9 CLUSTERS AND THEIR EUI DISTRIBUTION RANGE COMPARISON WITH THAT OF WHOLE DATASET

The graph (FIGURE 25) above shows the EUI distribution ranges in a cluster and can be compared to the whole building dataset. The EUI distribution for the cluster selected for the further analysis (cluster 1) is 132 kBtu/ft^2/year.similarly, the distributions for the

other 8 cluster can be clearly evaluated from the graph.

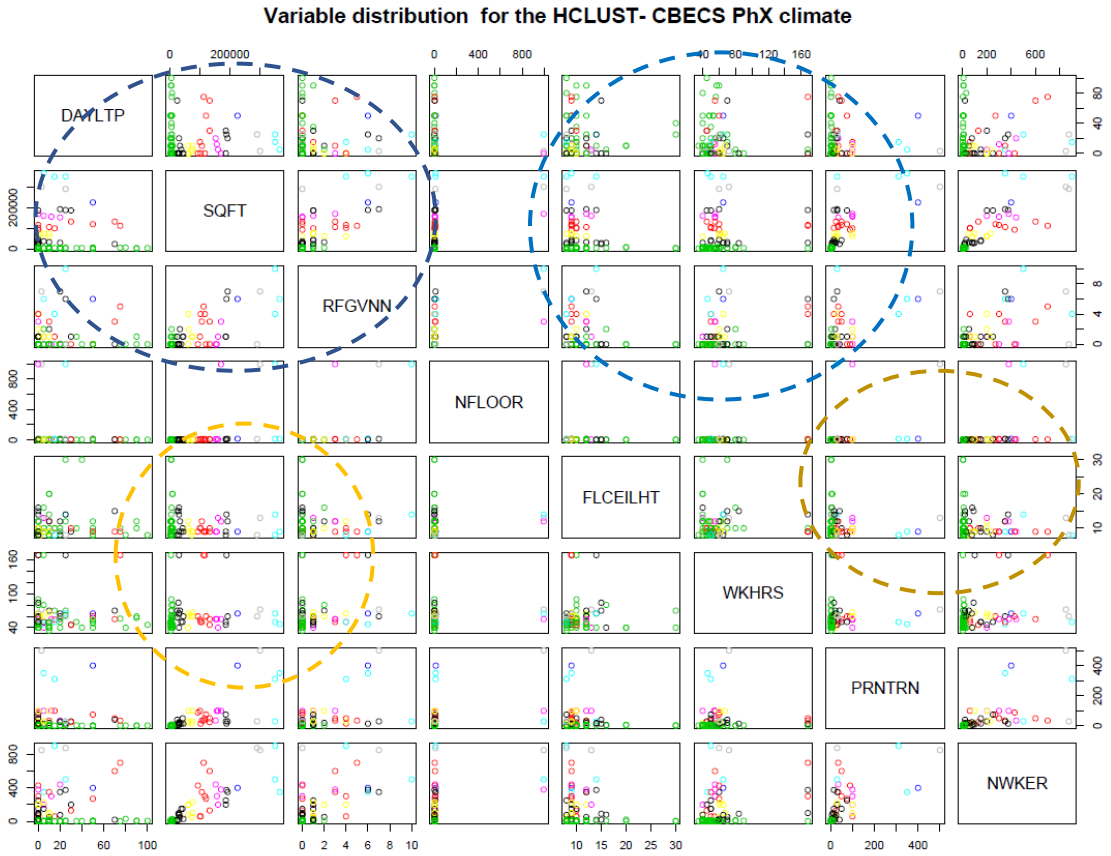

FIGURE 26: VARIABLE DISTRIBUTION PLOT FOR THE CLUSTERS: CBECS PHX CLIMATE

Based on these 9 clusters the CBECS dataset's importance variable on which the clustering was based can be analyzed from the variability distribution graph shown in above FIGURE 26. Each color represents a cluster from 1 to 9. The four-encircled region shows the effective demarcation of each cluster and its key deterministic characteristics. The clusters on both tails of the scatterplot region will effectively show the extreme distribution of the response variable.

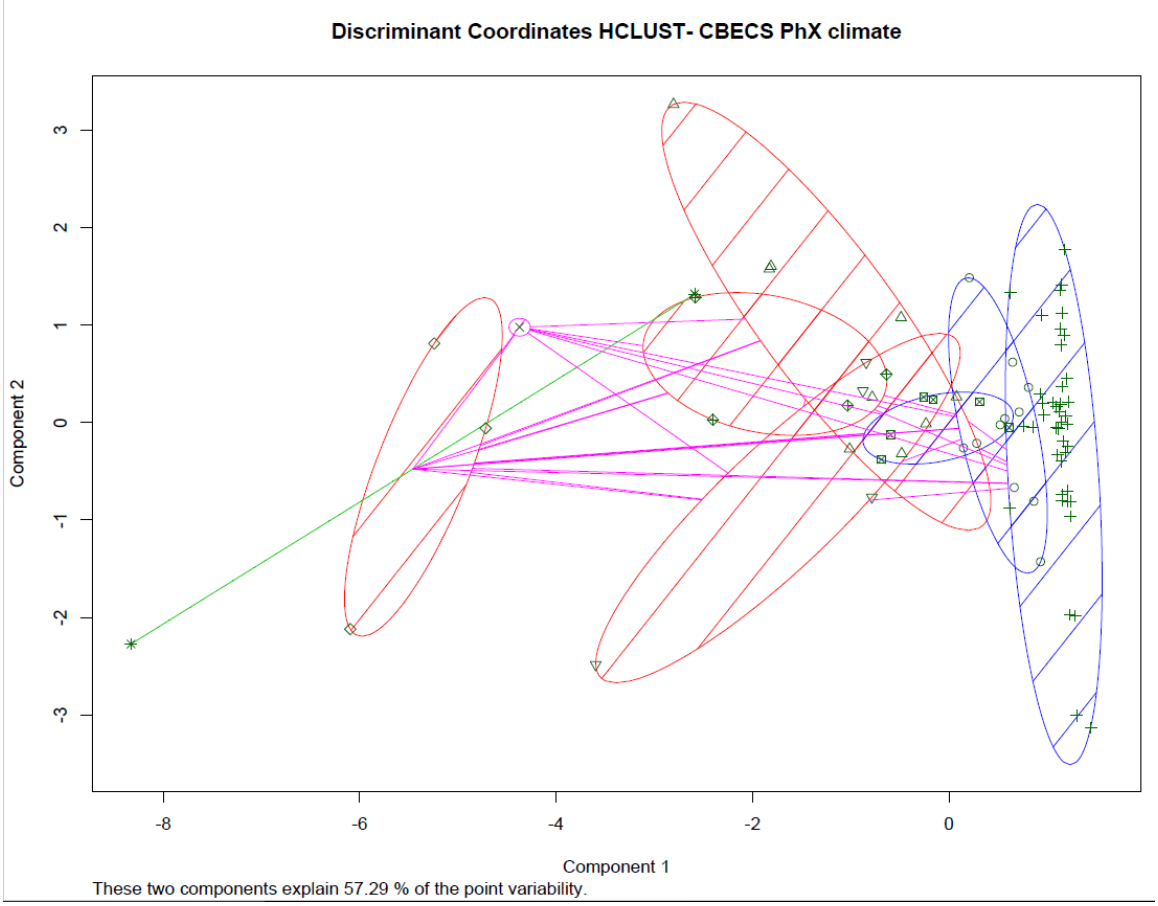

FIGURE 27: DISCRIMINANT PLOT FOR EACH CLUSTER

The FIGURE 27 shows the discriminant plot for the clusters formed on the selected building dataset. These Discriminant coordinates displays the primary differences between clusters, and is similar to principal components analysis. It defines the (dis)similarity between clusters as the pair-wise distance between all the respective centroids of each clusters.

The table lists the key characteristics of each cluster. The cluster number 6 and 7 should be reanalyzed for the anomaly in the number of floors value. Except this anomaly, all the clusters are distinctively identified by the initial analysis of the mediods.

TABLE 4: CLUSTER MEDIODS AFTER CONDUCTING THE HIERARCHICAL CLUSTERING ON SELECT CBECS DATA SET -PHX DATASET

Cluster mediods	1	2	3	4	5	6	7	8	9
DAYLTP	12.00	50	85	8	10	3	12.5	75	15
NFLOOR	3.00	8	2	4	10	994	994	4	11
FLCEILHT	11.00	9	10	11	12	13	13	9	8
WKHRS	53.00	63	54	168	50	72	60	168	50
SERVERN	3.00	23	3.5	6	18	65	17	136	50
COPIERN	4.00	50	5	5	100	60	30	9	46
RFGRSN	2.00	41	4	7	6	0	20	7	15

TABLE 5: VARIABLE DETAILS OF CLUSTER SELECTED FOR METHOD VALIDATION AND MONTE- CARLO SIMULATION

object id	BLDSHP	SQFT/floor	SQFT	YRCON	DAYLTP	NFLOOR	FLCEILHT	WINTYP	WKHRS	NWKER	PCTERMN	PRNTRN	SERVERN	COPIERN	RFGRSN	RFGVNN	EUI_total
1	1	15000	30000	1972	0	2	14	1	65	19	26	20	10	2	3	0	63.89
39	2	33000	33000	1967	10	1	12	2	70	53	49	20	1	0	2	2	75.83
45	2	15000	30000	1985	0	2	9	1	168	100	100	10	5	0	4	0	30.03
50	6	12000	24000	1999	0	2	8	3	58	14	25	6	4	2	1	1	46.24
53	2	10500	42000	1972	0	4	10	1	50	150	142	50	0	4	1	1	89.35
54	9	10000	30000	1952	0	3	16	1	40	80	30	6	0	6	0	0	62.11
57	2	37500	37500	1995	5	1	9	3	50	150	150	15	0	8	2	0	52.3
61	11	14000	28000	1985	0	2	15	1	84	91	2	15	0	5	8	0	22.04
70	6	12000	24000	1984	70	2	15	1	70	20	20	41	2	7	0	0	22.88
75	9	15000	30000	1973	20	2	8	3	65	77	35	25	1	0	2	0	13.63
78	6	14850	44500	1986	0	3	9	1	60	30	30	15	2	5	3	1	48.68

As defined in the previous chapter, each building is taken for the validation of the methodology of the simulation and a baseline model is created using the unique details extracted from the CBECS dataset and remaining inputs set from the ASHRAE 90.1 2010 compliant reference building. The details of these are provided in the

APPENDIX A. The findings from the simulations conducted on the building samples of cluster 1 and the validation of the proposed methodology by Monte Carlo simulations, are discussed next. The whole building simulations include individual building simulations of each 11 buildings and the same for the representative mediod building.

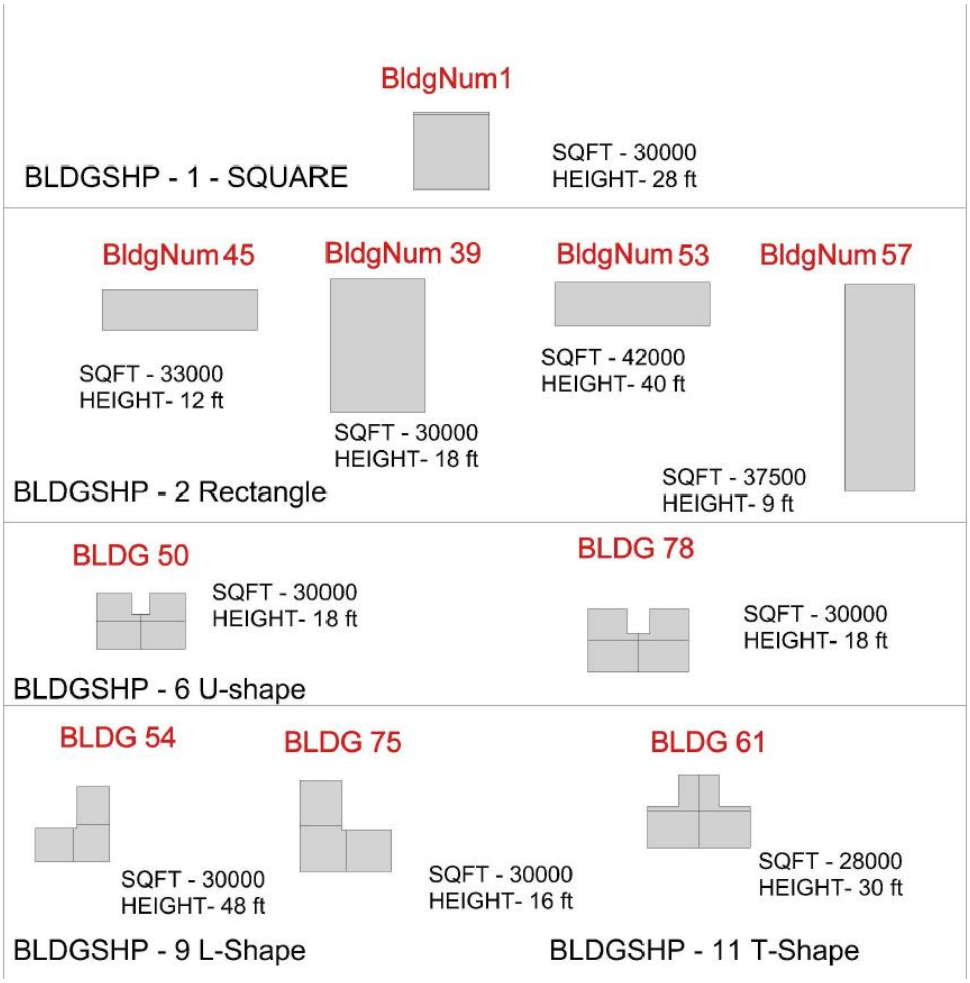

FIGURE 28: THE BUILDING GEOMETRY OF CLUSTER 1

4.3 Building Energy Simulation Analysis

1. Based on the preliminary information available pertaining to individual geometry of each building, the geometries were using the Rhino-Grasshopper scripts. The plan of

the resulting cluster is shown in FIGURE 28 below. Based on these geometries the scripting language was further utilized for the creation of zoning from the massing. The glazing, internal loads and schedules were also set similarly. For creating a common simulation conditions other than the key characteristics an ASHRAE 90.1 compliance model for each building was created. The variations of the EUI for the CBECS data and compliance model shown below.

1. Setting up baseline model allows the study team to conduct the uncertainty and sensitivity analysis. The variability towards each key parameter which are essential towards identifying feasible retrofits are identified in the previous section. The possible number of combination between these parameters will be in millions and thus not advisable to conduct the full variability analysis. Instead, the effective Latin hypercube sampling is done on the variable distributions. Based on this hypothesis a sample of 1000

different combinations are set up and batch simulation of each building model is carried out. The sample analysis of these batch simulations is explained below.

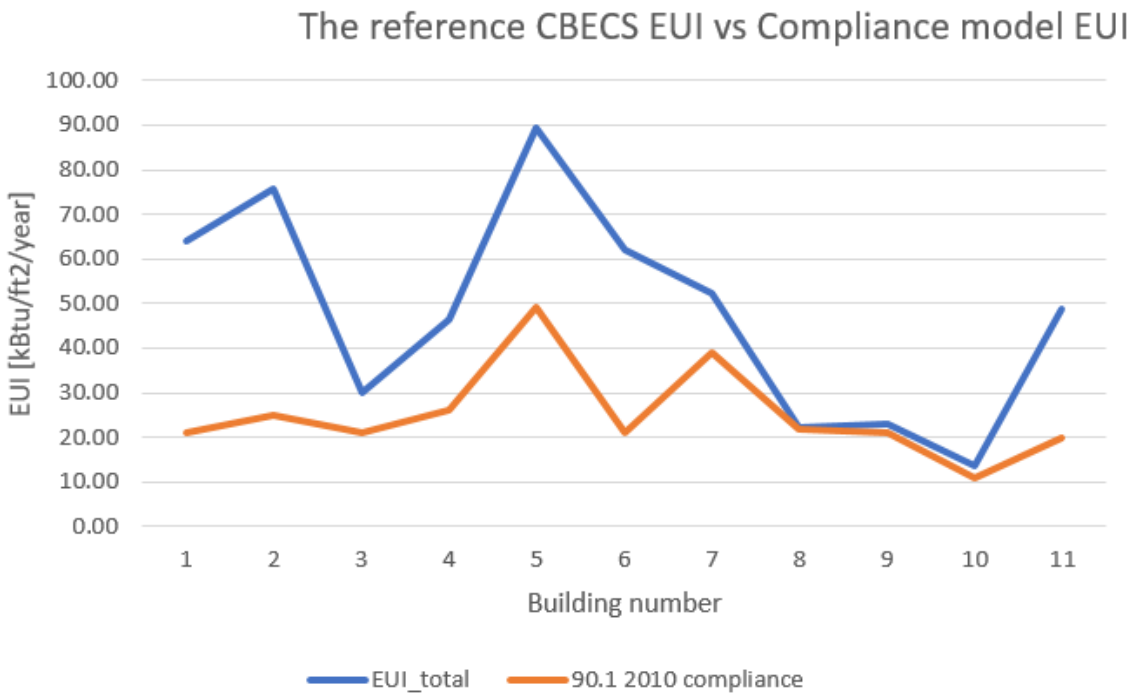

FIGURE 29: THE REFERENCE VS MODELED EUI

a. Uncertainty Analysis

The following FIGURE 30 shows the distribution of the EUI as a response variable for the Monte-Carlo Simulation conducted on the compliance model of the building number 45 of the cluster 1. The EUI distribution varies from as low as 7 kbtu/ft2/year to 68 kbtu/ft2/year. The red dashed line shows the geometric mean of the distribution. 10 similar graphs for the remaining buildings were also obtain following this procedure. Since, the uncertainty analysis was conducted based on the variability of the 13 operational variables (listed before), the variability in the EUI of the building is understandably due to the combined impact of these variables. There have been studies

which have focused on the impact of an individual parameters on the EUI (or any other response variable that is under study) and singular variables impact is then added up and the combined impact of the parameters as a whole is estimated. These approach is limited application to only the variables that are mutually independent i.e. effect of one variables is not correlated with the effect of another variable on the response variable. Since, building's thermal behavior is a combined effect of multiple correlated variables and their variability the uncertainty analysis study is helpful in establishing the combined impact of these variables. Thus, the uncertainty analysis also helps to study the variability in EUI distribution of each building individually as well as can be compared with the EUI variability of the representative mediod building of the cluster. In short, the uncertainty analysis of individual buildings would lead to parameters screening for simplification of the simulation model, evaluating the robustness of the simulation model and thus effectively provides direction for the retrofit soluitons. The results of this are listed in the next chapter.

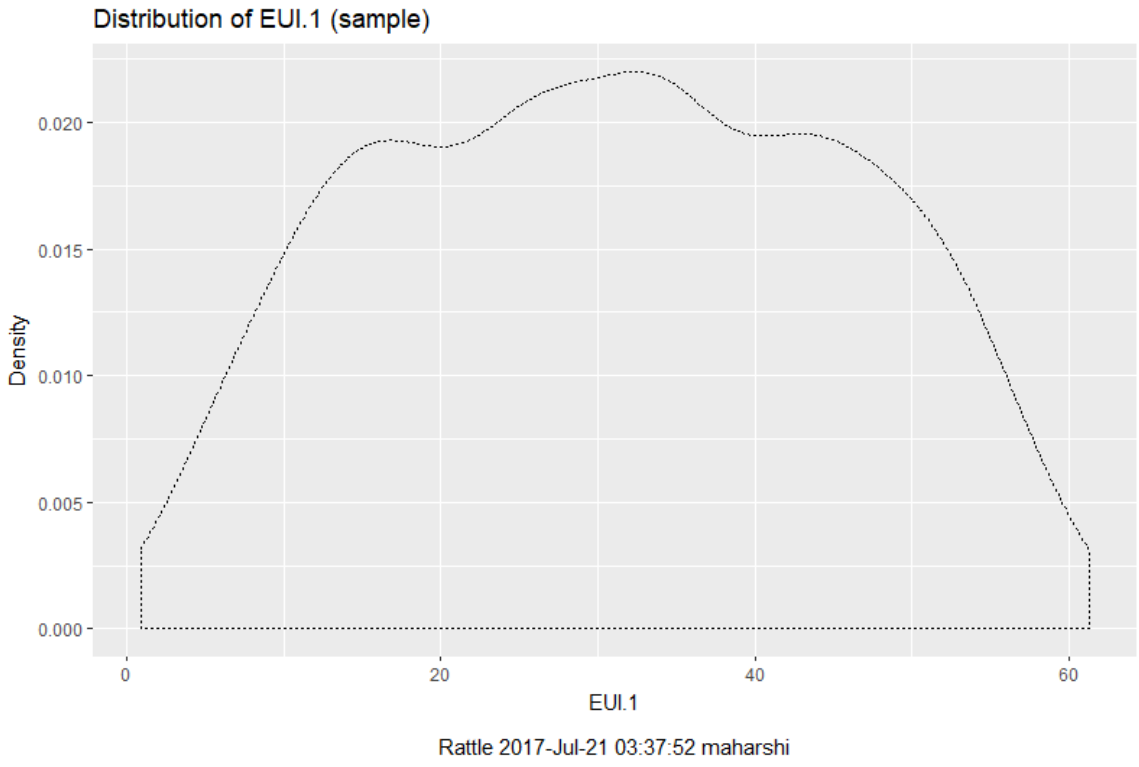

FIGURE 30: BUILDING #1 EUI HISTOGRAM -MONTE CARLO SIMULATION – UNCERTAINTY ANALYSIS CAN BE USED FOR IDENTIFYING THE EUI VARIABILITY DUE TO INPUT PARAMETERS IMPACT ON THE SAME AND THE EUI VARIABILITY OF INDIVIDUAL BUILDINGS AND THE MEDIOD BUILDING

b. Sensitivity analysis

Since, there are there are millions of possible combinations which are responsible for affecting the variability in the EUI – response variable due to the 13 variables at hand, understanding various scenarios of EUI variations due to input parameter combinations can be a very confusing for the users with no or very little expertise of the data science principles. Here, advanced techniques of data visualization can help explain this in a more concise and visually effective way. Here, in the present study the sensitivity analysis has been studied and tried to analyze various retrofit scenarios with the help of parallel coordinates plot for the 13 input variables and 1 response variable i.e. EUI. The theory of how a parallel coordinate plot works is explained in the previous chapter. Let's

understand and analyze this parallel coordinate plot for the sample building number 1 and 1000 whole building simulations' results obtained for the 13 variables' 1000 different combination from their respective distributions. Although, a parallel coordinate plot can be confusing at first sight (FIGURE 31), especially given its limited use in the building science community, they can often be quite rich on closer inspection. To make more the visualization more informative, let's remove certain input parameters based on their Pearson coefficient (commonly referred as Pearson R test) value, which is an indicator of the sensitivity of each input variable towards the target variable i.e. EUI. The Pearson coefficient of each variable as compared to the EUI is given below.

TABLE 6 SENSITIVITY ANALYSIS OF THE VARIABLES ON THE BATCH SIMULATION RESULTS - PEARSON COEFFICIENT VALUE FOR EACH VARIABLE

Sensitivity analysis on batch simulation					
Input Variable	EUI - sensitivity	Input Variable	EUI - sensitivity	Input Variable	EUI - sensitivity
Wall Conduction	0.008	Roof Emissivity	-0.010		
Wall Absorption	-0.043	Window Conduction	0.014	LPD	-0.057
Wall Emissivity	-0.022	Win-ST	-0.035	Heat-Setpt	-0.003
Roof-Cond	-0.016	Infiltration	0.031	Cool-Setpt	-0.044
Roof-Abs	0.069	EPD	0.005		

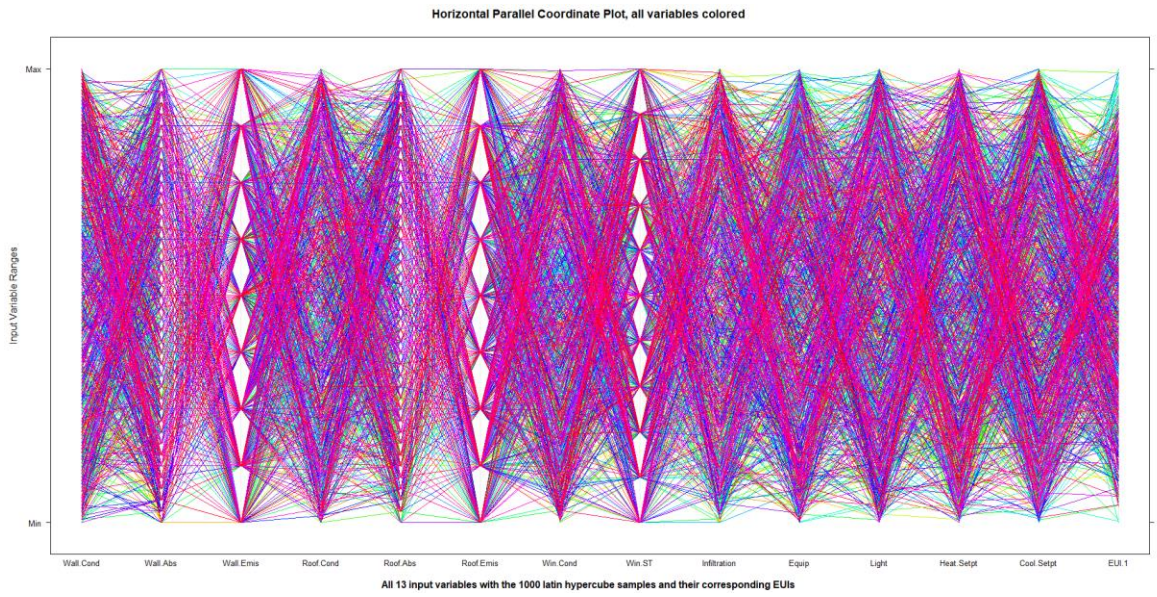

FIGURE 31 PARALLEL COORDINATE PLOT FOR ALL 13 VARIABLES VIS-A-VIS EUI

To make this graph more specific let's analyze the variables with positive and negative sensitivity towards EUI separately, reduce number of samples to 100 and remove the colors from the plot. Here, notice that the roof absorption and how leaky the envelope is showing the maximum sensitivity towards the energy usage intensity. This behavior can be described as a perimeter dominant energy loads which is usually the case for the phoenix climate office buildings. This also sets the potential opportunity for focusing the retrofits scenarios in a way that leads towards energy efficient envelope strategies.

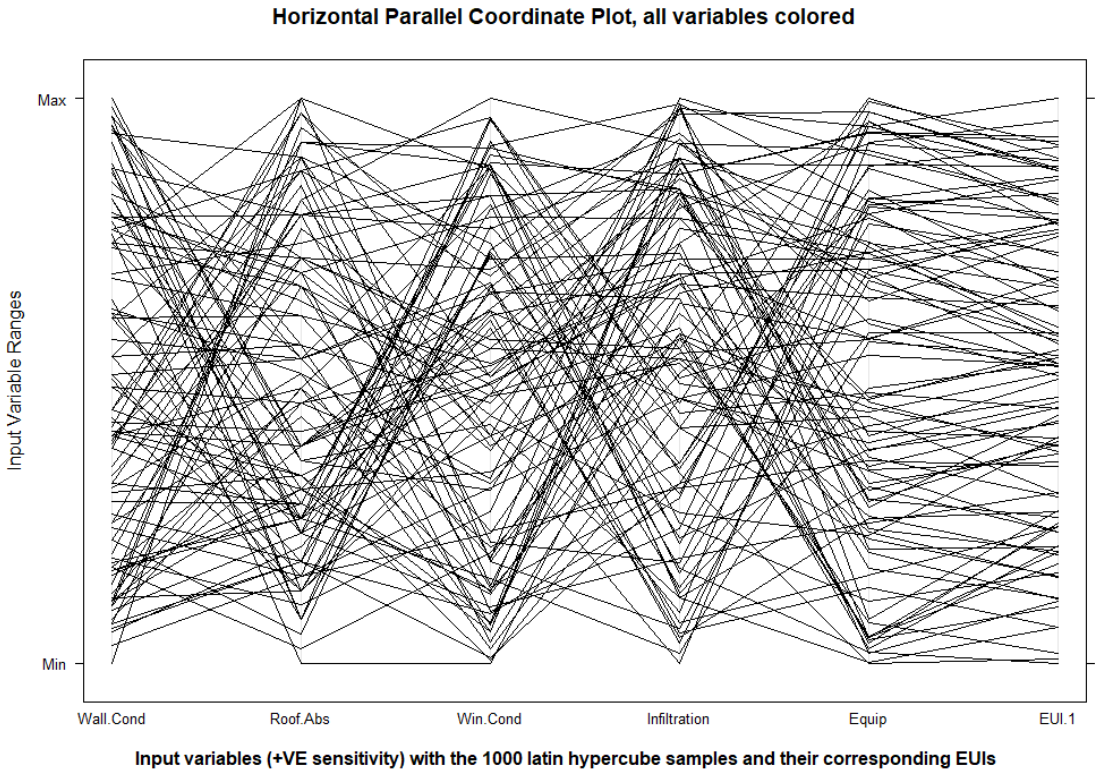

FIGURE 32 5 DISTINCT VARIBLES WITH POSITIVE CORRELATION WITH EUI, ALL 1000 SIMULATIONS,

Here, let's understand the graph shown in FIGURE 31. It stimulates visual description of the sensitivity of each variable towards the energy use intensity. The envelope related parameters show strong correlation between them and EUI individually whereas the equipment loads being manly the part of the core loads (when making the perimeter vs core load comparisons). Now, having established that the perimeter loads are to be reduced let's further analyze the effect of the increased resistance (reduce conductivity)

of the wall material i.e. adding more insulation and how effective it is on energy intensity.

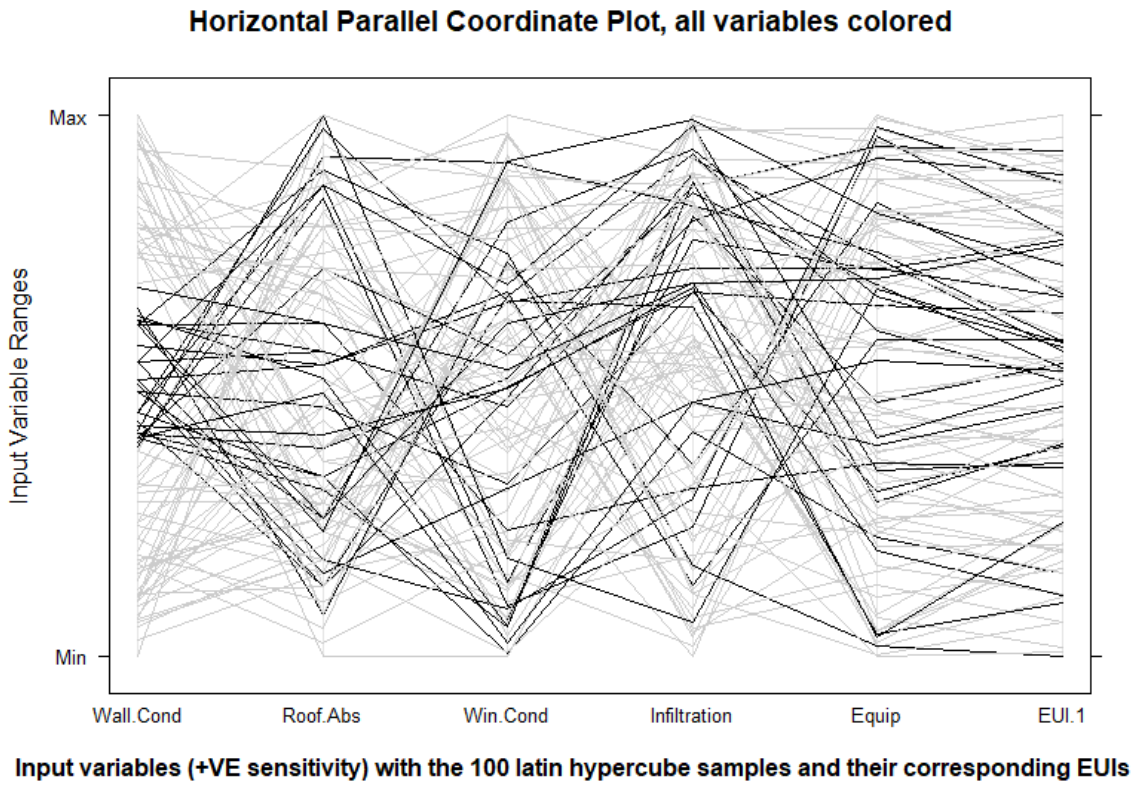

*FIGURE 33 COMPARING THE EUI DISTRIBUTION FOR THE SELECT 5 VARIABLES WITH FOCUS ON THE WALL CONDUCTIVITY OF 0.3 W/M*K TO 0.6 W/M*K, THE GROUPINGS FOR THIS ARE VERY COMPELLING AS THE EUI VALUES ARE DISTRIBUTED ACROSS SPECTRUM, THUS ALONG WITH LOW WALL RESISTANCE OTHER PARAMETERS SHOULD ALSO BE CONSIDERED TO DERIVE BETTER EUI TARGETS.*

The FIGURE 33 above shows the EUI distributions of the building 1 for the wall conductivity values in the range of 0.7 $W/m^2/K$ to 1.1 $W/m^2/K$ (ASHRAE 90.1 2010 compliance value and 50% more effective conductance than the compliance value). It is evident from the uniform distribution of the EUI that just effective wall conductance is not sufficient in achieving the better groupings of the resultant EUI of the selected building. Other envelope measures should also be added and the combined impact of the same should be studied for effective target achievement. After adding up all the envelope related variables with their value ranging from the ASHRAE 90.1 2010 compliance value

to 50% more effective value, the EUI of the resulting combinations can be grouped as per the FIGURE 34 given below. It is evident that as from the graph's red lines is that EUI reduction targets may be achieved by various combination of affecting input parameters but if the sensitivity of these input parameters is studied then the most effective combinations out of all possibilities can be identified with the help of what-if analysis as explained here.

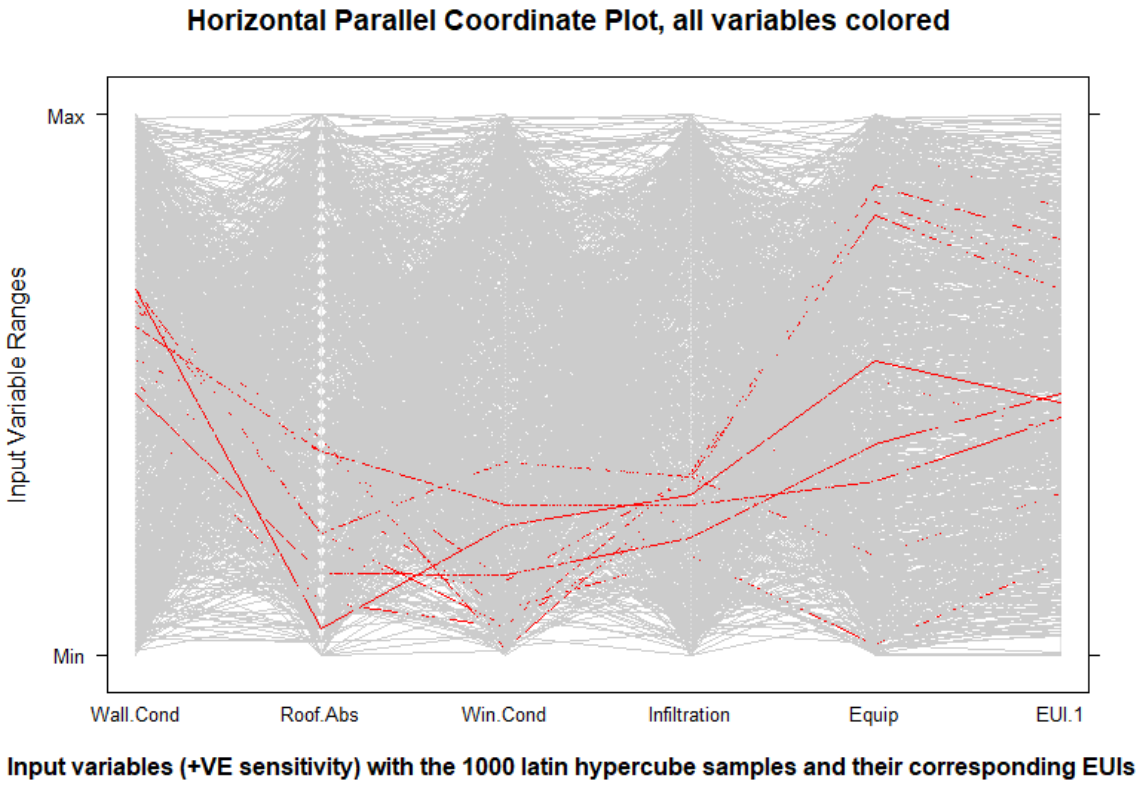

FIGURE 34 FINAL SENSITIVITY ANALYSIS RESULTS FOR IDENTIFYING THE MOST EFFECTIVE COMBINATION OF THE ENVELOPE IMPROVEMENT STRATEGIES OUT OF THE DISTINCT 1000 VARIED COMBINATIONS WITH EFFECTIVE DATA VISUALIZATION TECHNIQUE AND WHAT-IF ANALYSIS

5. RESULTS AND CONCLUSION

The results of this research assemble several key findings of the proposed methodology of prototype definition used for the purpose of conducting urban scale energy analysis.

Machine Learning Algorithms:

a. Random forest: While the conventional statistical analysis concepts of linear regression failed to explain the relationship between the independent variables and response variable i.e. energy use intensity of the large number of buildings studied, the unsupervised ensemble learning algorithm explains the relationship effectively. Unlike, OLS models, the random forest algorithms consider categorical variables on as is and trains the model. This is key in building energy analysis studies which is an amalgamation of several continous and categorical variables. The variable importance, thus, identified in the regression model provides strong foundation for carrying out the clustering.

b. Clustering: Since, hierarchical clustering algorithm is based on dissimilarity distance between the studied parameters, it helped the research aim of dividing the database into effective number of clusters which are not based on user's preference of number of clusters but on the contrary, it provides the information on basis of which the clustering algorithms divided the cluster of 81 varied buildings into 9 unique clusters. This in turn enabled the researcher team to establish parameters considered for the retrofits recommendation.

c. The mean behavioral of buildings of cluster 1 vis-à-vis the behavior of the mediod building found by the clustering: The complete data set of the building is clustered into 9 distinct clusters. The EUI distributions of each one of them is plotted with respect to the EUI distribution of the whole sample size in the FIGURE 35 CLUSTERS AND THEIR % EUI DISTRIBUTION WRT EUI DISTRIBUTION OF THE WHOLE DATA SET below.

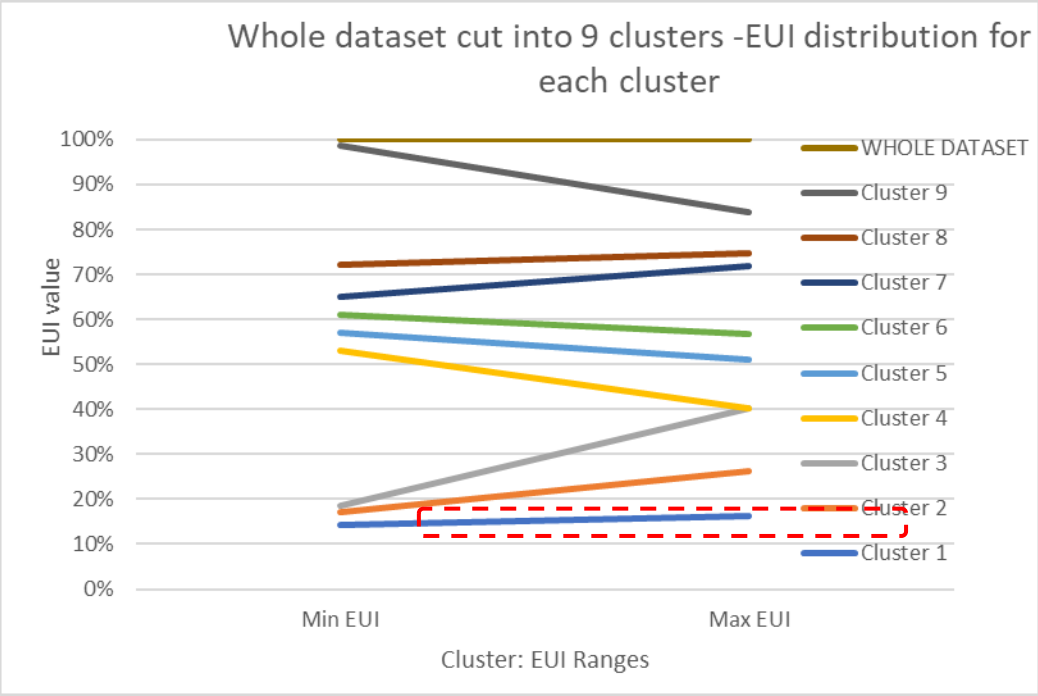

FIGURE 35 CLUSTERS AND THEIR % EUI DISTRIBUTION WRT EUI DISTRIBUTION OF THE WHOLE DATA SET

The baseline models of the selected 11 buildings of the cluster 1 are processed for batch simulations and results of these simulations were established. The descriptive analytics of the selected cluster's important variable which drives the clustering algorithm are given in the *TABLE 1: BUILDING PARAMETERS USED FOR THE UNCERTAINTY ANALYSISTABLE 1* below.

TABLE 7 DESCRIPTIVE ANALYTICS OF THE SELECTED IMPORTANT VARIABLES OF CLUSTER1

Descriptive Analytics of the IMP VAR -Cluster #1						
	DAYLTP	WKHRS	PRNTRN	SERVERN	COPIERN	RFGVNN
Mean	9.55	70.91	20.27	2.27	3.55	0.45
Standard Error	6.34	10.35	4.19	0.93	0.88	0.21
Median	0.00	65.00	15.00	1.00	4.00	0.00
Mode	0.00	65.00	15.00	0.00	0.00	0.00
Standard Deviation	21.03	34.33	13.90	3.07	2.91	0.69
Sample Variance	442.27	1178.49	193.22	9.42	8.47	0.47
Kurtosis	8.42	7.74	1.05	3.57	-1.44	0.98
Skewness	2.84	2.62	1.27	1.84	0.05	1.32
Range	70.00	128.00	44.00	10.00	8.00	2.00
Minimum	0.00	40.00	6.00	0.00	0.00	0.00
Maximum	70.00	168.00	50.00	10.00	8.00	2.00
Sum	105.00	780.00	223.00	25.00	39.00	5.00
Count	11.00	11.00	11.00	11.00	11.00	11.00

The simulation results: Based on the set LHS of size 1000, the whole building energy simulation of each building is plotted below FIGURE 36. The representative mediod building's energy performance was also studied. The EUI distributions of the same are plotted in FIGURE 37 below. The mean distribution of the individual buildings is in the range of 5 to 60 kBtu/ft2/year whereas the mediod of the building's EUI distribution is varied in a larger range of 20 to 120 kBtu/ft2/year.

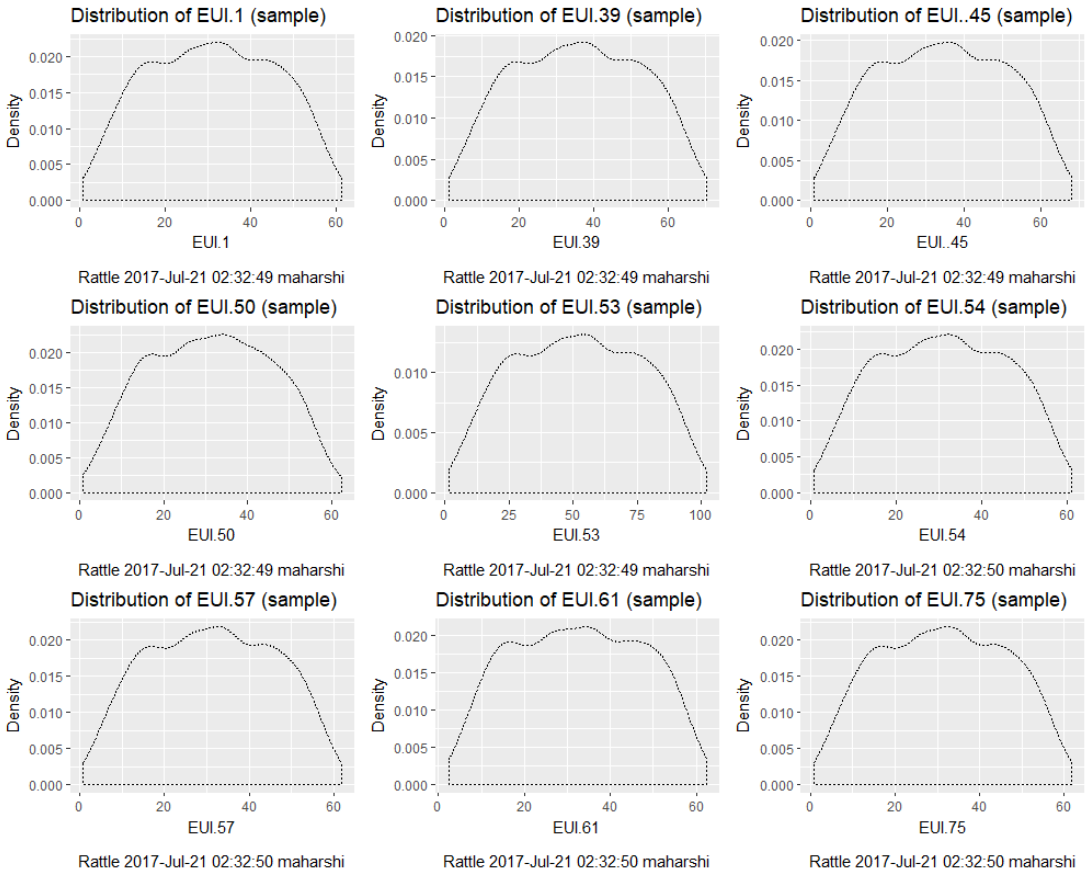

*FIGURE 36 EUI DISTRIBUTIONS OF EACH BUILDING'S 1000 SIMULATIONS: THE PROBABILITY DISTRIBUTION FUNCTIONS OF EACH BUILDINGS' EUI ARE SHOWN, THE PDFS OF MOST OF THE BUILDINGS OF THE CLUSTER ARE RANGING IN VALUE OF EUI 0 BTU/FT^2*H/YEAR TO 60 BTU/FT^2*H/YEAR AND THE DISTRIBUTION IS ALSO IDENTICAL.*

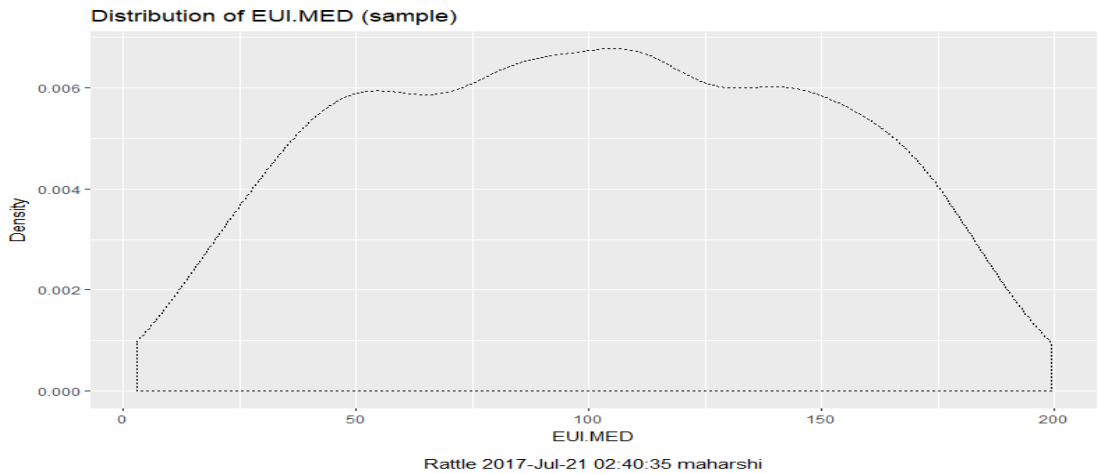

*FIGURE 37 EUI DISTRIBUTIONS OF THE REPRESENTATIVE MEDIOD BUILDING'S 1000 SIMULATIONS: THE PROBABILITY DISTRIBUTION FUNCTIONS OF BUILDINGS' EUI ARE SHOWN, THE PDFS OF MOST OF THE BUILDINGS OF THE CLUSTER ARE RANGING IN VALUE OF EUI 0 BTU/FT^2*H/YEAR TO 200 BTU/FT^2*H/YEAR*

AND THE DISTRIBUTION IS CONSIDERED TO BE REPRESENTATIVE OF THE 11 ABOVE BUILDINGS' EUI PDFS.

REFERENCES

Addison, M.S. (1988). A Multiple Criteria Satisficing Methodology For the Design of Energy-Efficient Buildings, Masters' Thesis, Arizona State University, Tempe.

Andrienko, N., & Andrienko, G. (2006). Exploratory analysis of spatial and temporal data: a systematic approach. Berlin: Springer.

Bae, C. (2008). *Analyzing random forests.* (Unpublished Doctor of Philosophy). UNIVERSITY OF CALIFORNIA, BERKELEY, Berkeley.

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York, NY: Springer Verlag.

Breiman, L. (2001). Machine Learning,45(3), 261-277. doi:10.1023/a:1017934522171

Cerezo, Carlos et al. "Three Methods for Characterizing Building Archetypes in Urban Energy Simulation. A Case Study in Kuwait City" Proceedings of BS2015: 14th Conference of International Building Performance Simulation Association, Hyderabad, India,7-9 December 2015

Dutta, R. (2013). A Visual Analytics Based Decision Support Methodology For Evaluating Low Building Design Alternatives, Masters' Thesis, Arizona State University, Tempe.

Ewen, R. B. (1971). Introduction to review of inferential statistics. Introductory Statistics for the Behavioral Sciences,145-155. doi:10.1016/b978-0-12-245050-1.50024-9

Everitt, Everitt, Brian, & MyiLibrary. (2011). Cluster analysis (5th ed., Wiley series in probability and statistics). Chichester, West Sussex, U.K.: Wiley.

Guha, S., Rastogi, R., & Shim, K. (1998). Cure. ACM SIGMOD Record,27(2), 73-84. doi:10.1145/276305.276312

Hansen, P., & Jaumard, B. (1997). Cluster analysis and mathematical programming. Mathematical Programming,79(1-3), 191-215. doi:10.1007/bf02614317

Hierarchical Clustering. (n.d.). Clustering,31-62. doi:10.1002/9780470382776.ch3

Jibonananda, S., & New, J. (2014). Oak Ridge Institutional Cluster Autotune Test Drive Report. doi:10.2172/1146987

Keirstead, J., Jennings, M., & Sivakumar, A. (2012). A review of urban energy system models: Approaches, challenges and opportunities. Renewable and Sustainable Energy Reviews,16(6), 3847-3866. doi: 10.1016/j.rser.2012.02.047

Kim, M., Lee, H. W., Dou, W., & Jung, W. (2009). An Analysis of Wind Field around the Air Quality Monitoring Station in the Urban Area by Using the Envi-met Model. Journal of the Environmental Sciences,18(9), 941-952. doi:10.5322/jes.2009.18.9.941

Korolija, I., Marjanovic-Halburd, L., Zhang, Y., & Hanby, V. I. (2013). UK office buildings archetypal model as methodological approach in development of regression models for predicting building energy consumption from heating and cooling demands. Energy and Buildings,60, 152-162. doi: 10.1016/j.enbuild.2012.12.032

Massaro, J. M. (2005). Clustering, Single Linkage. Encyclopedia of Biostatistics. doi:10.1002/0470011815.b2a13087

Parker, C. T., Taylor, D., & Garrity, G. M. (2010). Exemplar Abstract for Actinomadura glomerata Itoh et al. 1996 and Actinocorallia glomerata (Itoh et al. 1996) Zhang et al. 2001. The NamesforLife Abstracts. doi:10.1601/ex.7594

Rapid energy modeling - Autodesk. (n.d.). Retrieved July 17, 2017, from http://sustainability.autodesk.com/available-solutions/rapid-energy-modeling

Reference Buildings by Building Type: Medium office. (n.d.). Retrieved July 18, 2017, from https://energy.gov/eere/downloads/reference-buildings-building-type-medium-office

Reinhart, C. F., & Davila, C. C. (2016). Urban building energy modeling – A review of a nascent field. Building and Environment,97, 196-202. doi: 10.1016/j.buildenv.2015.12.001

Sarle, W. S., Jain, A. K., & Dubes, R. C. (1990). Algorithms for Clustering Data. Technometrics,32(2), 227. doi:10.2307/1268876

Theodoridis, S., & Koutroumbas, K. (2009). Clustering Algorithms II: Hierarchical Algorithms. Pattern Recognition,653-700. doi:10.1016/b978-1-59749-272-0.50015-3Snyder, S.C., Reddy, T.A., & Addison, M.S. (2013). Automated Design of Buildings: Need, Conceptual Approach, and Illustrative Example, ASHRAE Transactions, Volume 119, Issue 1, p1.

Tursilowati, L., Sumantyo, J., Kuze, H., & Adiningsih, E. (2012). The integrated WRF/Urban modeling system and its application to monitoring urban heat island in Jakarta-Indonesia. Journal of Urban and Environmental Engineering,6(1), 1-9. doi:10.4090/juee. 2012.v6n1.001009.

APPENDIX A

EDA OF THE CBECS SELECT VARIABLES

a) Linear/Non-linear Relationship Amongst the Design Variable to EUI

For each variable in the select dataset of CBECS, the relationship with the target variable needs to be studied. The following graphs shows how design variables and operational variables are related to the target response EUI. The data is color-coded based on its climate zone, i.e. variable “PUBCLIM”. The graphs on the diagonal shows each variable distribution. To study this graph, each column represents the relationship of the corresponding diagonal member of the column and other parameters. Starting with (FIGURE 38) square footage of the building, the graph shows that there is no direct linear relationship with the energy usage and how the floor area varies with each building. Same is the case for year of construction and heating degree days. The EUI is proportional to the day lit percentage, number of floors in the building and floor to floor height. From the other two graphs (FIGURE 39,FIGURE 40) operational variables and their relationship with the EUI is established. As expected, the rise in number of equipment tends towards the increased energy usage. Similarly, occupancy factors also show the same tendency.

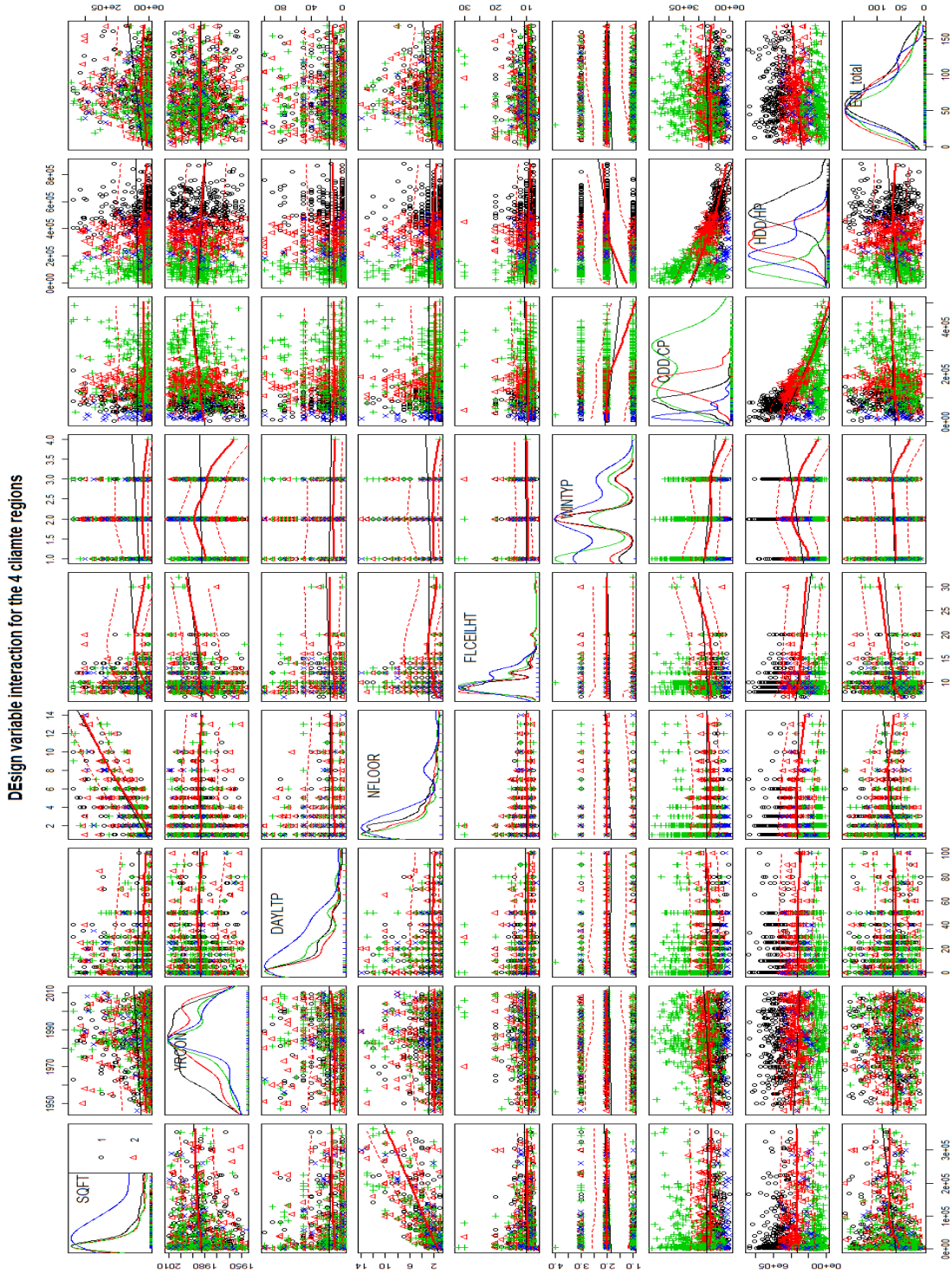

FIGURE 38: SELECT DESIGN VARIABLES' INTERACTION WITH THE RESPONSE VARIABLE - EUI

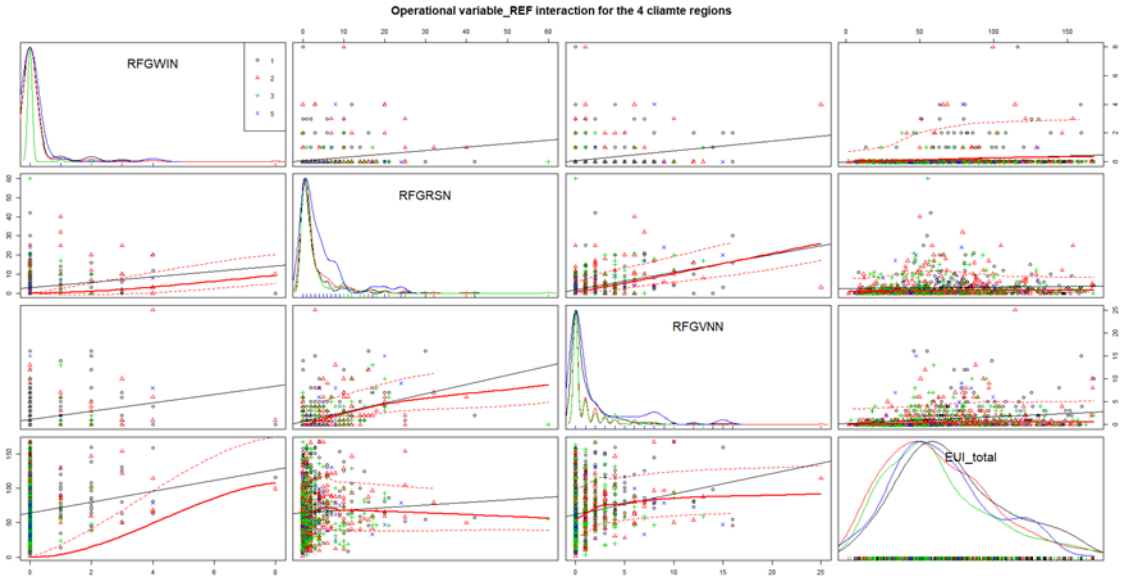

FIGURE 39: SELECT OPERATIONAL VARIABLES (-1) ' INTERACTION WITH EUI

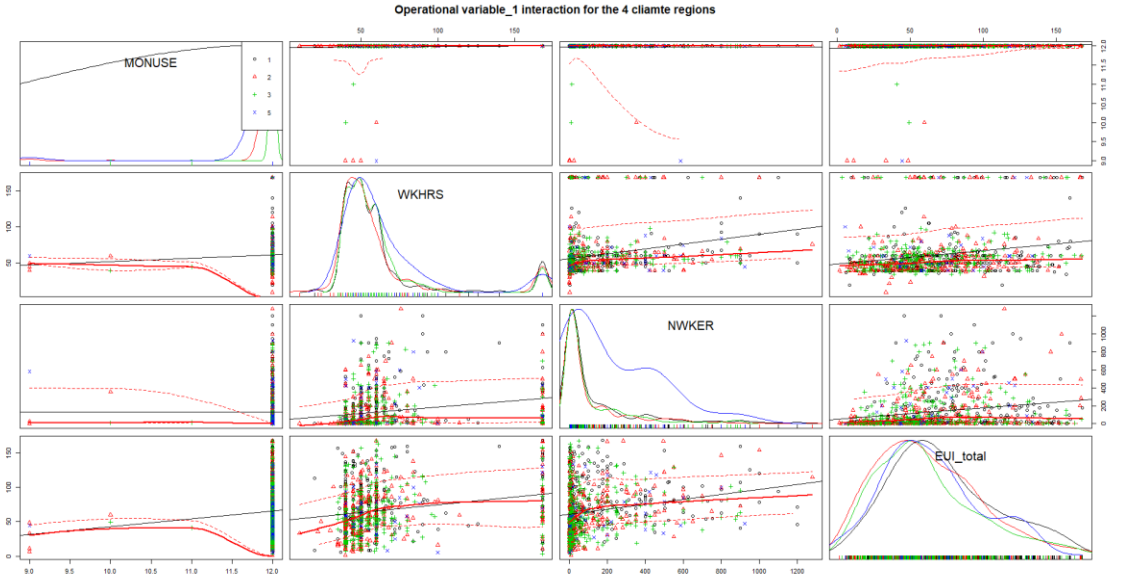

FIGURE 40: SELECT OPERATIONAL VARIABLES (-2) ' INTERACTION WITH EUI

b) Correlation Between Variable

The building design variables,

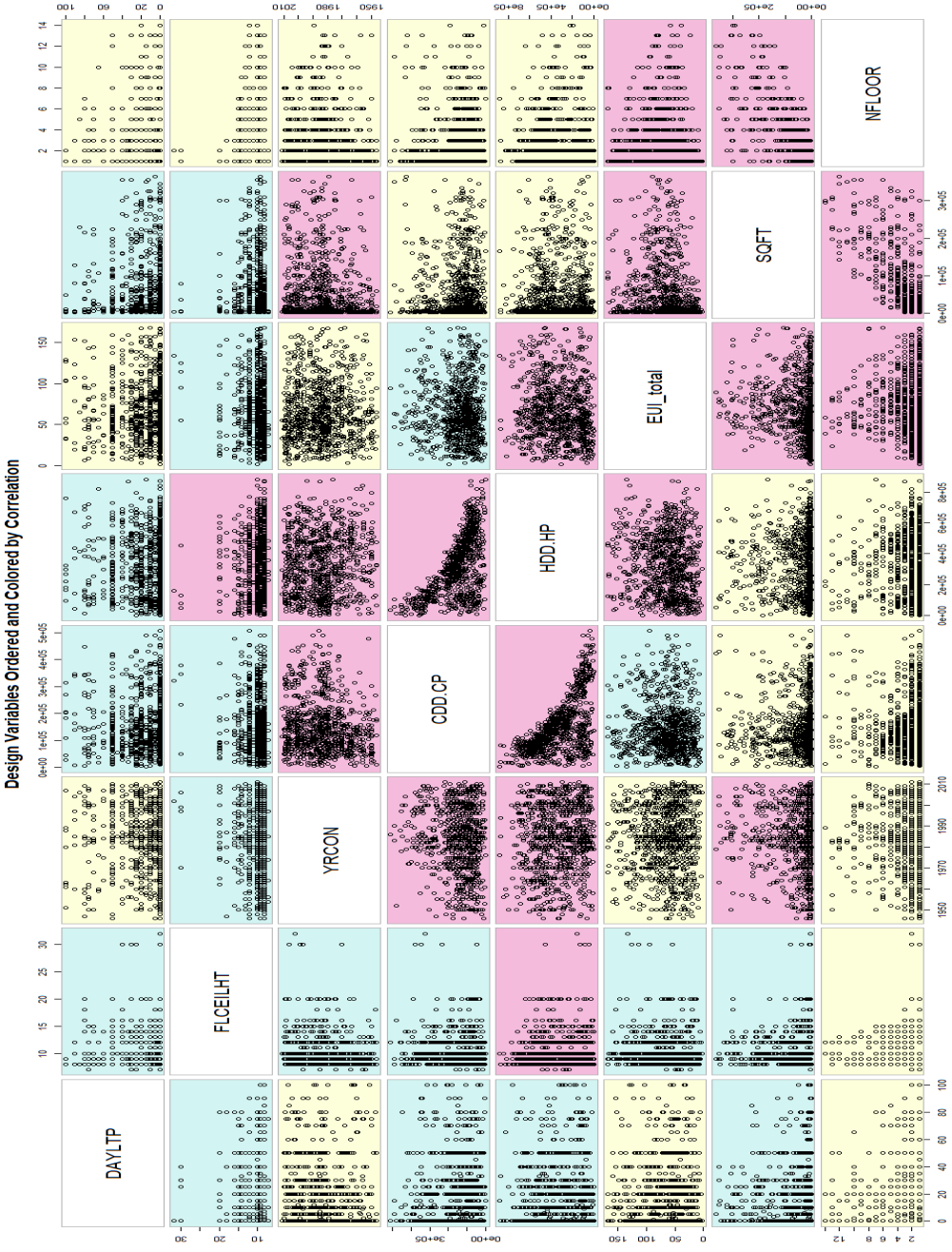

FIGURE 41 DESIGN VARIABLES - CORRELATION BY DATA MINING

FIGURE 42 OPERATIONAL VARIABLE- PLUG LOAD - CORRELATION MATRIX

operational variables i.e. occupancy, plug and process load variables and the energy usage pattern are studied to understand the correlation with the help of data mining techniques. The gradient of correlation varies across columns in a way that highly correlated variable set closest to the diagonal. Thus, highest correlation is observed at the lower right corner.

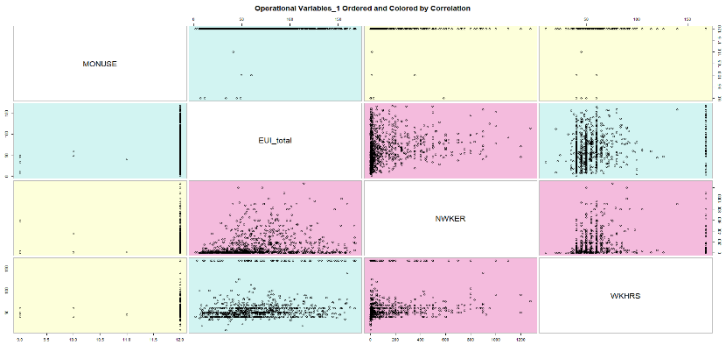

FIGURE 43 OPERATIONAL VARIABLE-OCCUPANCY LOAD - CORRELATION MATRIX

As it is evident from the building design variables, the energy usage is least correlated floor height and cooling degree days but highly correlated towards heating degree days and square footage of the building and number of floors in a building. The year the building was built are neutral towards the energy usage of the building, i.e. energy usage is not directly correlated to the year of construction.

Similarly, the occupancy variable and equipment variable were studied. As shown in FIGURE 42 and FIGURE 43**Error! Reference source not found.** the energy usage is most correlated to the number of occupants and number of occupied hours the building. In case of the plug loads, the number of computers are the most correlated to energy usage patterns of the buildings in the database.

APPENDIX B

PNNL PROTOTYPE MEDIUM SIZE OFFICE BUILDING SPECIFICATIONS

- Building model specification for PNNL prototype medium size office building located in the hot and dry climate of Phoenix (Climate zone 2B)

FORM	
Total Floor Area	53,600sq.ft. (4982 sq.m)
Dimensions	163.8ft. x 109.2ft. (49.926 x 33.28 m)
Number of Floors	3
Floor to floor height	13ft. (3.96m)
Thermal Zoning	Each floor has four perimeter zones and one core zone. Perimeter 40%, Core 60% of floor area. Perimeter zone depth: 15ft. (4.57m)
Window-to-Wall Ratio	33%
Window Locations	Evenly distributed along four façades
Glazing sill height	3.4ft. (1.02m)
Glazing height	4.4ft. (1.31m)
ARCHITECTURE	
Exterior wall construction-	Steel-Frame Walls (2X4 16IN OC): with wood siding, wall Insulation+ 1/2 in. gypsum board
U-factor/ R-value	0.0677 Btu / h * ft2 * °F (R-14.775)
Roof Construction-	Built-up roof: roof membrane+ roof insulation+ metal decking
U-factor/ R-value	0.049 Btu / h * ft2 * °F / R-20
Window-	Metal framing- double pane, 0% Operable area
U-value	0.417 Btu / h * ft2 * °F (2.369 W/meter*K)
SHGC/ Visible transmittance	0.8/ 0.89
Foundation Type	Slab-on-grade floor (unheated): 8" concrete slab
Internal Floor	4" in concrete with tiles
Interior Partitions	Air wall
Air Barrier System	0.43 cfm/sf (0.0002 m^3/sec/sq.m)
Infiltration	

BUILDING HVAC SYSTEM & CONTROLS	
System Type	Ideal Air Load System
Thermostat Setpoint	Cooling -75°F (24°C) / Heating- 70°F (21°C)
Thermostat Setback	Cooling -80°F (26.7C) / Heating- 60°F (15.6 C)
Supply air temperature	Maximum 104°F (40°C), Minimum 55°F (12.8°C)
Ventilation per area	2.1426 cfm/sq.ft. (0.00043m^3/sec/m^2)
Ventilation per person	17 cfm/person (0.008m^3/sec/person)
BUILDING OPERATIONS & INTERNAL LOADING	
Occupancy Schedule	7am - 5pm Mon-Fri
Lighting & Equipment Schedule	7am - 5pm Mon-Fri
Fan Schedule	7a-10p WD, 7a-6p Sat, Sun off
Lighting power density	1 W/sq. ft. (10.76 W/sq.m)
Receptacle Load (W/sf)	0.75 W/sq. ft. (8.07 W/sq.m)
OCCUPANCY	
Average people	268 persons
No of person/sq.m area	0.0538