City-scale single family residential building energy consumption prediction using Genetic Algorithm-Based Numerical Moment Matching Technique

Elham Jahani[a], Kristen Cetin [b], In Ho Cho [a]

[a] Iowa State University, Department of Civil, Construction and Environmental Engineering, 813 Bissell Rd, Ames, IA 50011

[b] Michigan State University, Department of Civil and Environmental Engineering, 428 S. Shaw Lane, East Lansing, MI 48824; cetinkri@msu.edu

Abstract

Growing energy consumption in urban areas has increased the importance of planning for future energy systems. Thus, improving the modeling abilities for predicting energy consumption at the city scale is critical. In this study, a Genetic Algorithm-Based Numerical Moment Matching (GA-NMM) method is adopted as a primary uncertainty estimation technique to predict the electricity consumption of a large dataset of single family homes by utilizing key features in energy audit and assessors data. This data is used as an input to the GA-NMM to develop a set of index buildings and associated weighting factors that represent statistical characteristics of the dataset. Energy models are then developed for the index buildings using physics-based energy modeling in EnergyPlus. These, in combination, are used to estimate the energy behavior of single family homes of the studied dataset.

The proposed method is applied to a large dataset of 8,370 single family homes in Cedar Falls, Iowa, where the expected annual and monthly electricity consumption from the model is calculated and compared with measured data. The expected site electricity consumption for single family buildings in Cedar Falls is estimated as 10,219 kWh/yr, which is within 6% of the measured average annual electricity consumption. At a monthly level, the Coefficient of Variation of Root Mean Square Error and Mean Bias Error are 7.8% and 4.5%, respectively. This method can be used to generate small set of representative homes for demonstrating the energy behavior of a larger set of homes.

Keywords:

City-scale energy modeling, Genetic Algorithm-Based Numerical Moment Matching, Index buildings, single family homes, residential buildings

1 Introduction

The large majority of energy modeling efforts of buildings has generally focused on models of energy consumption of single buildings, where both physics-based [1–3] and data-driven [4–6] models have been extensively developed for this purpose, each with its own advantages and disadvantages depending on the intended application. Given the diversity of building types and uses, highly varied occupant behaviors particularly in residential buildings [7], and significant ranges in age and efficiency of the building stock in the U.S., it is understandable that this single building focus is among the more common modeling goals. However, in recent years there has also been a focus on the development of methods to predict energy consumption for clusters of buildings, both at the neighborhood [8] and at the city scale [9–11], as well as the regional and national scale [12,13]. These methods are highly useful for city- and utility-level planning purposes.

Multi-building modeling methods generally fall into two categories, including bottom-up and top-down modeling, as discussed and reviewed in several recent review papers [12,14–16]. Top-down approaches utilize the aggregated energy consumption data of the building stock, and relate the total energy consumption of the region with variables such as gross domestic product (GDP), employment rates, and price indices, climatic conditions, and housing construction and/or demolition rates [12,14]. Bottom-up approaches [15,16] account for the energy consumption of individual buildings or groups of buildings, then extrapolate the modeled sample to represented region. This approach can generally be accomplished in two different ways, using statistical methods (SMs) or engineering methods (EMs). SMs rely on data-driven approaches such as regression analysis of historical data to relate building energy consumption to influential parameters that impact energy consumption. There are several techniques that can be categorized under this umbrella. Different studies have reviewed the advantages and disadvantages of these techniques. For example in a study by Do and Cetin [5], six data driven techniques such as Change-Point Models [17,18], Artificial Neural Network [19,20], Genetic Programing [21], Bayesian Network [22], Gaussian Mixture Models [23], and Support Vector Machines [24] are compared, focusing on pros and cons of each technique specifically for residential building energy modeling. In this study it is concluded that the complexity of prediction models and the amount of input data needed significantly varies between methods, and the tradeoffs between more complex as compared to less computationally complex methods vary [5]. In another study done by Swan and Ugursal

[14], three techniques classified as SMs, including Neural Network (NN) [25,26], Conditional Demand Analysis (CDA) [27,28] and regression methods [29,30], are evaluated. The positive aspects of SMs techniques discussed include previous research demonstrating the use of such techniques for modeling of occupant behavior, end-use energy consumption, and macroeconomic and socioeconomic effects, whereas the negative attributes include reliance on historical consumption data and the necessity of a large sample to model a variety of the characteristics [14,31].

For Ems, the energy consumption of homes is determined based on physical phenomena, including heat transfer and thermodynamic relationships. Common input data to physical models include building properties such as geometry, envelope fabric, equipment and appliances, climate properties, as well as indoor environmental conditions, occupancy schedules and equipment use. There are several techniques that can be used to model building energy consumption at different scales. These can be categorized into the following methods: (a) distribution method, (b) archetype technique, and (c) sample techniques [14]. The distribution method uses the distribution of the number of energy-consuming devices in each building, the common rates for energy consumption on a device-level basis, typical profiles of consumption of these end-uses, and their efficiency to calculate total energy consumption [32,33]. In this method the interrelationship between the use patterns of the energy-consuming devices typically have not been taken into account [14]. For the archetype technique, also sometimes called the prototype technique, the housing stock is broadly classified into a set of buildings according to size, building type, or other characteristics. The energy consumption of each modeled archetype buildings is used to represent the population of buildings classified under that archetype, then scaled up to represent the share of that archetype building in the region of study [14,34,35].

Finally, for the sample technique, an actual sample of building data is used as the input into the model. The captured sample data can represent the wide variety of buildings in different regions. If the sample is representative of the regional or national building stock, using weighting factors, the energy consumption of the studied population can be estimated [14,36]. The number of samples depends the sampling method. For example, Monte Carlo (MC) simulation [37,38] generates a large number of samples. This can be computationally demanding and time consuming. Moreover, the probability density functions (PDF) of some building characteristics may not follow a convenient parametric distribution [39]. Another sampling method, Mean and Sigma (MS) [40] requires a smaller sample size compared to MC. In MS, the number of samples exponentially increase based on the number of predictor variables.

Using the methods mentioned above, most multi-building energy modeling efforts to date have focused on representing the characteristics and/or energy consumption of the residential and/or commercial building sectors. Given that residential and commercial buildings represent nearly 75% of the U.S. electricity consumption [41], and approximately 40% of energy consumption [42], it is understandable why these two building types are the main focus of most research efforts. Due to the fundamental differences in the use and energy demand profiles of residential and commercial buildings, however, often the models for energy consumption of these two main building types are separated.

Larger scale data collection and modeling efforts to characterize the U.S. building stock include the RECS [43] and CBECS [44] datasets for residential and commercial buildings, respectively. These statistical datasets are collected every few years by the U.S. government from a statistically representative sample of buildings across the country. They include data on annual energy consumption and a variety of energy-related building characteristics. These two datasets are highly referenced, and are often considered as datasets by which others compare their results. Other recent national government-funded efforts include the development of ResStock [45] for residential and ComStock [46] for commercial buildings for analysis of the building stock characteristics and energy consumption at the state and national level, and UrbanOpt at the district or neighborhood level [47]. ResStock, similar to the others, utilizes over ten different public and private datasets as outlined in [45], latin-hypercube sampling to represent the 80+ million single family homes in the U.S. with a representative sample of homes using 6,000 conditional probability distributions, and physics-based energy models using EnergyPlus run using supercomputing resources. It is meant to be used to identify the energy efficiency measures that will save the most energy and money across the building stock. These methods, however, are not developed for city-level energy analysis, as the level of granularity of the data is not sufficient to be city-specific.

Closer to the city-scale, UrbanOpt [47], as well as other similar GUI-based modeling tool development efforts such as UMI (urban modeling interface) [48], enables the modeling of all buildings in a city district individually, as opposed to a statically representative set of buildings. These methods require detailed GIS-based information on the building floor areas and locations as well as the categorization of each building into a prototype building (e.g. small office, restaurant, retail, etc.), which each have a set of standard energy consumption-related assumptions. These models benefit from the use of highly detailed information, however, given the significant effort needed for energy modeling at the building-level, constructing an energy model building-by-building of a large city with potentially hundreds of thousands of buildings, is not highly practical in most applications, with current technologies. In addition, for many cites, the level of detail of meaningful building characteristic information can be limited to assessor's data, unless there is a city or local mandate requiring more energy-related information to be reported. This mandate exists in some U.S. cities (e.g. Energy Conservation Audit and Disclosure in Austin, TX [49]), but is more commonly required for public buildings. Lack of detailed data for individual buildings is one of the main challenges. There are some efforts to expand building-level energy characteristics of city building stock, such as those discussed in [5], however this is not the reality for many cities currently.

Numerical Moment Matching (NMM) is a method of representing a large population with a substantially smaller sample size while preserving statistical moments (e.g., mean, variance, skewness, kurtosis, etc.). NMM's sample size is order of magnitude smaller than that of other random sampling methods such as MC and MS. However, to date this method has not been used for representing the building stock and its energy-related characteristics. For accuracy comparison, a study by Cho and Porter [39] on large-scale earthquake risk assessment, examined the difference between the moments of the sample and surveyed population (i.e. error) using the NMM and MS methods. Results showed that in the MS method this error is 1.62% in calculating the first moment of the sample $E[x]$, and for the higher moments up to $E[x^5]$, the error increased to up to 28%.

However, for the NMM method, the maximum error for $E[x^5]$ was found to be 0.005% for the same variable. This indicates that error of the MS method is reliable only in calculating mean value for a symmetric and regular form of data and performs poorly in estimation of higher-order moments for each predictor/variable. Therefore, when the original data distribution is considerably irregular or when the higher-order moments matter, NMM performs better compared to MS method.

The initial version of NMM was based on the multivariate Newton-Raphson (mNR) scheme, which sometimes causes numerical divergence and initial-value dependency [50]. To overcome these limitations, a Generic Algorithm (GA) has been coupled with mNR scheme in the context of NMM (denoted as GA-NMM) to exhibit no restrictions to irregular distributions, large sizes, or many variables of engineering data [50].

The objective of this study is to utilize GA-NMM to develop a set of statistically representative index buildings for a larger building population to evaluate electricity consumption. The ability to generate a small set of characteristic buildings for use in analysis of the larger dataset's energy behavior reduces subsequent analyses, statistical implications, and computational intensity. Given the fundamental differences in energy behaviors of residential and commercial buildings, this research specifically focuses on a dataset of residential, single family homes. Building energy audit data for Cedar Falls, Iowa, is used as a case study to evaluate the performance of the proposed technique in predicting the energy behavior of a large population of houses using data from those homes sampled. The results of this study are compared with measured monthly electricity consumption data to validate the applied methodology. As compared to using building characteristics and energy-related assumptions from some of the larger residential building datasets currently available (e.g. ResStock [45] and RECS [43]), developing localized parameters to use for model development helps to decrease the computational costs while maintaining accuracy of the results.

2 Datasets

Four main datasets are used in this study, all of which are utilized to characterize the homes in the Cedar Falls region, and for the development and validation of the GA-NMM model. These are described as follows. In this section, we also benchmark these homes' characteristics as compared to homes in the Iowa region, and the U.S. to enhance the understanding of the applicability of the methods proposed herein.

The first dataset is publicly-available assessor's data, used for tax purposes, for the residential buildings in Black Hawk County, Iowa, where Cedar Falls is located. This is used for characterizing the overall housing stock in Cedar Falls. Throughout the U.S., basic information about the residential building stock is maintained as public records, including building area, age, and type (single family, duplex, townhouse, etc.), number of bedrooms, and heating, ventilation, and air conditioning (HVAC) system, among others. Based on this dataset, the median age and size (total living area) of homes in Cedar Falls are 61 years old, and 109 m^2, respectively. The U.S. building stock (2017) is a median of 42 years in age and 139 m^2 in area [51]. In the West North Central Region of the U.S., which includes Iowa, homes are a median of 44 years of age, and 148 m^2. Thus, the homes in Cedar Falls studied herein are slightly older and smaller in comparison to the broader region in which they are located.

The second dataset includes energy audit data for a subset of the residential buildings in the Cedar Falls region. This contains field-measured energy audit data for 531 single family homes, four variables of which were targeted in this study for use as inputs into the GA-NMM method. These include building area, attic insulation R-value, window type and cooling system efficiency (SEER – seasonal energy efficiency ratio), as reported in Figure 1. These variables were chosen for two reasons. First is based on the results of a one-at-a-time (OAT) sensitivity analysis of the impacts of these characteristics on electricity consumption of residential buildings under typical climate conditions in Iowa, as discussed in [52]. In comparison to other variables included in the energy audits, these factors have a higher impact on electricity consumption. This is further discussed in the method section. These were also chosen as they can be considered independent variables, which is necessary for the proposed GA-NMM method. A correlation analysis indicates that these variables have correlation coefficients of less than 0.23.

For this data, given that the homes included in the energy audit data are a subset of the homes in Cedar Falls, first their characteristics are compared to the overall building stock in Cedar Falls from the assessor's data. The energy audit homes overall have a similar age distribution to those in Cedar Falls (Figure 1). In addition, the average age of the homes in both datasets is also similar (within 2%). The energy audit homes, however, are slightly larger than the Cedar Falls residential building stock, with 71% being less than 139 m^2 as compared to 59% in the audit data. Given these characteristics, while the audit characteristics of all the homes in Cedar Falls are not measured, the size and age generally follow similar trends.

Compared to the building characteristics in the U.S. and in the state of Iowa from [45], as shown in Figure 1, the homes in Cedar Falls generally have a similar distribution of attic insulation values (Figure 1b). The distribution of homes with different window types and cooling efficiency values is generally fairly uniform across the surveyed homes in Cedar Falls. Thus in this dataset, there is a slightly higher portion of homes that have low-e windows and SEER 15 or higher efficiency systems as compared to the U.S. and Iowa.

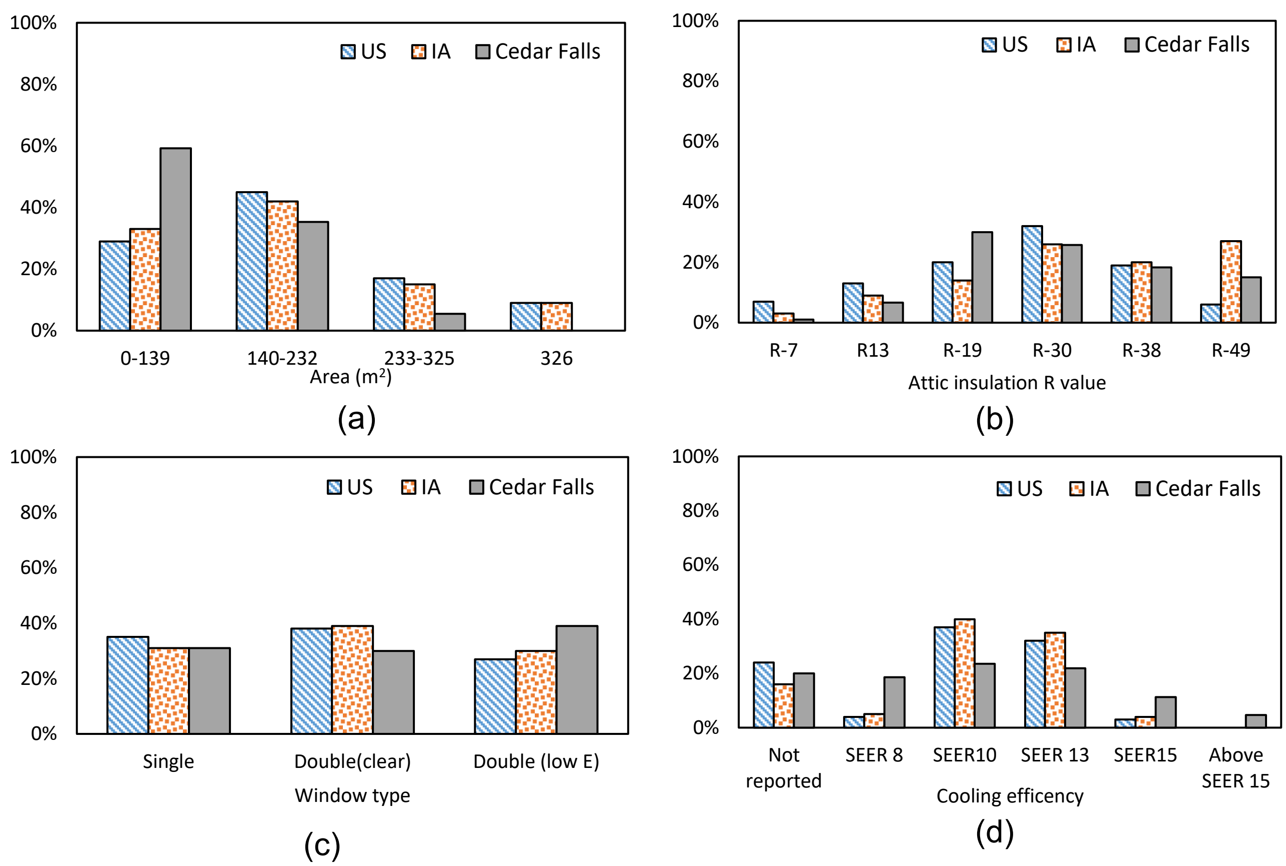

Figure 1: Characteristics of the Cedar Falls residential building stock; a) building area, b) Attic insulation R value, c) window type, d) Cooling systems efficiency, as compared to the U.S. and Iowa [45].

The third dataset is monthly measured electricity consumption of single family homes obtained in collaboration with Cedar Falls Utilities. This dataset is used to validate the applied methodology with measured data. This dataset includes 101,220 electric bills for 8,370 single family homes from January to December 2010. On average the annual electricity consumption of the homes in this dataset is 10,872 kWh. In comparison to the averages reported in RECS (2015) in the U.S. and Midwest, of 10,726 kWh and 10,051 kWh, respectively, these values are highly similar. We also compare the average electricity consumption of homes in Cedar Falls, to the electricity consumption of the homes included in the energy audit data. The energy audit data homes annually consume 10,968 kWh on average, which is 1% higher than the average annual use of overall Cedar Falls building stock, indicating that the audit dataset generally follows the same energy use trends as that of the larger Cedar Falls buildings stock. It also should be noted that the electricity consumption data of the homes used for model development (energy audit data of 531 homes), is not used for model verification. Thus, only data not used to develop the model (i.e. out-of-sample data) is used to assess the model's performance.

The fourth dataset is measured weather data for Cedar Falls in the year 2010. This dataset is obtained from a database that includes over 10,000 stations, collected since 2001, with sufficient detail to create hourly weather files for use in energy modeling methods [53]. Cedar Falls in Iowa is located in a cool-humid climate according to ASHRAE Climate Zone definitions. In this climate zone the Heating Degree Day (HDD) per year are in the range of 3000 to 4000 [54].

3 Methodology

The methodology described herein includes two main steps. First is the development of a model to predict the annual and monthly site electricity consumption for single family homes in Cedar Falls. This is accomplished through the utilization of the GA-NMM method to obtain a set of index buildings and associated weighting criteria which statistically represent the energy related features of the larger dataset of buildings, which are used to develop building energy models of the index buildings, then used to predict the annual and monthly site electricity consumption of the building population. The second step is the verification of model results with measured monthly electric use.

3.1 Predicting electricity consumption for single family buildings

3.1.1 Genetic Algorithm-Based Numerical Moment Matching (GA-NMM)

The NMM technique is a point estimation method that is applicable for large scale data by generating a small set of representative samples of a population. Different techniques have been presented in the literature [39,50,55] that aim to enable fast and stable moment matching. The more traditional NMM method based on the mNR has disadvantages such as severe numerical divergence and initial-value dependency. However, these limitations have been addressed by Karr et al. [56] and Cho et al. [50], by stabilizing NMM with a genetic algorithm (GA), hereafter designated as GA-NMM. GA-NMM has no restrictions for irregular distributions, large sizes, or number of variables. In this method, it is assumed that an independent variable *X* describes the attribute(s) or feature(s) of the studied population. It is also assumed that a function $Y = g(X)$ can relate the response variable *Y* to the independent variable *X*. The goal is to estimate expectation ($E[Y]$), and variance ($Var[Y]$) of *Y*. In this study building electricity consumption is considered as the resultant parameter which is a function of building characteristics, including cooling system efficiency, building area, window type and attic insulation R-value. It should be noted that independency of the variables is the only requirement for this method, which is met for all the proposed variables.

After checking interdependency of the variables, the next step is to generate a discrete probability function that can statistically represent a continuous probability distribution function (PDF) of the studied variable. This discrete probability can be considered as probability mass function (PMF) which consists of positions and their associated weights for each variable. In this study, as recommended in [39], to avoid computational complications, the PMF for each variable consists of three positions and their associated weights. The weights and positions of the PMF are chosen such that first five moments (mean, variance, skewness, etc.) of both the discrete PMF and the original PDF are identical. More detailed information on how to obtain the three positions and their weights for each variables in a way that the resulted PMF exactly match the first five moments of variable's distribution is given in [39]. The next step combines all PMFs and develops a junction distribution of all the variables' PMF. The final result is *2n+1* index buildings where *n* is the number of variables that are selected as impactful variables on building electricity consumption. For example, if there are four key variables from the population, this would result in twelve positions and associated weights. Considering each variable as an axis with three positions and their associated weights, combining all 4 axes would result in a 4-dimensional junction distribution. All of these axes share one of their three positions to make the centroid of the 4-dimensional junction distribution. Therefore, the final number of positions in the 4-dimensional junction distribution would be 4 multiplied by 2 plus 1, which results in 9 index building. The weight for the center of these axis is obtained using, $w_1 = 1 - \sum_{2}^{2n+1} w_i$, where n is the number of variables and w_{i} is the associated weight of each index building. Further explanation on n-dimensional Moment Matching using PMFs with three positions is given in [39].

3.1.2 Utilization of the GA-NMM method for building energy prediction

To generate index buildings using the GA-NMM method, the independent key features that have significant impact on electricity consumption of the single family buildings must be identified. Moreover, these features should vary among the studied building population, and utilized data. For example, 93% of the surveyed buildings in Cedar Falls use gas as a heating fuel. Given the high percentage of homes with gas as a heating fuel, this variable, while it does impact the electricity consumption of a building, in this dataset is considered a single variate feature and is not included in the GA-NMM analysis. Among the main categories that are measured in the energy audit data, two of them (water heating and heating components) are nearly uniformly gas-based equipment in the dataset, and thus have almost no impact on electricity consumption for the studied homes. It is noted that for homes in other climate zones or regions of the U.S. this may not be the case, however, considering the referenced data, this level of uniformity does occur, and thus is treated as appropriate. This is also consistent with the RECS data on annual household site end-use consumption in very cold/cold regions in which the electricity consumption for space heating and water heating contributes to less than 8% of the total site energy consumption [57]. Considering these criteria, the key features that have higher impacts on electricity consumption of single family homes are chosen, including building area, cooling system efficiency, attic insulation R-value, and window type (single pane (clear), double pane (clear, metal frame, air filled), and double pane (low E, non-metal frame, air filled).

As explained in Section 3.1.2, after identifying impactful variables, the PMFs of each variable with three positions are generated using GA-NMM. Then by combining the resultant twelve PMFs and their associated weights, and generating a joint distribution, nine index buildings and their associated weights are developed. It should be noted that the obtained index buildings with their associated weights can statistically represent the expected electricity consumption of the population when the weighted average is calculated from all the index buildings. Therefore, an individual index building does not represent the characteristics of the population.

3.1.3 Modeling energy consumption for the index buildings

To model energy consumption for each resulting index building, a building energy model is developed in EnergyPlus [58], using the software BEopt (Building Energy Optimization) v.2.8.0.0 to establish the base models. This interface is specifically designed for implementing residential building energy models. Since weather data highly impacts the energy behavior of buildings, for modeling the energy behavior of the index buildings, measured weather data in 2010 for Cedar Falls, Iowa is used, paralleling the measured energy use data used during this same time period. The majority of the input parameters utilized in the building energy models originate from the assumptions and data discussed in the Building America House Simulation Protocol [60].

To better represent the homes included in the measured data, several factors are adjusted. The cooling and heating setpoints are modeled to be 22.7^{o}C, and 20^{o}C respectively. This is consistent with the housing characteristics and baseline consumption for U.S. single family homes in this region, as discussed in [45], where 89% of the Iowa single family homes have a cooling setpoint of 22.7^{o}C and 61% have a heating setpoint of 20°C [45]. In addition, the miscellaneous plug loads (MELs) are considered be double that of the value calculated by the empirical formula discussed in the Building America Housing Simulation Protocols [61]. This was adjusted since the energy use for non-HVAC loads for very cold/cold regions has been found to have the highest level of site energy consumption comparing to other climate regions [57]. All modifications are applied to all nine index building models.

The annual site electricity consumption is then calculated for each index building. To determine the expected value of annual site electricity consumption for the whole sample, obtain from the population, Equation (1) is utilized,

$$E[Y] = \int \mathrm{PDF}(\mathbf{X}) Y(\mathbf{X}) d\mathbf{X} \underbrace{\cong}_{\mathrm{NMM}} \sum_{i=1}^{2n+1} w_i Y_i \tag{1}$$

where Y_i is annual site electricity consumption for each index building and w_i is the associated weight used for each index building *i*, $\mathbf{X}$ is underlying variable vector that determines the outcome Y. As easily seen, the main role of NMM here is to replace the intractable integration of Equation (1) with the simple weighted summation of outcomes.

Therefore, the expected monthly and yearly energy consumption of single family homes for the studied population is predicted using the weighted average of all index buildings. To measure the confidence of the statistical conclusion of the expected value, the expected standard deviation (denoted as *STD*) is calculated using Equation (2). The $E[x] \pm STD$ is used to represent the statistical acceptable interval for the obtained result. The same procedures are applied for calculating expected electricity consumption for each month.

$$STD = \sqrt{E[Y^2] - E[Y]^2}$$

$$E[Y] = \sum_{i=1}^{2n+1} w_i Y_i \, ; \; E[Y^2] = \sum_{i=1}^{2n+1} w_i Y_i^2 \tag{2}$$

3.2 GA-NMM Model Validation

To validate results obtained from GA-NMM technique, monthly measured electricity consumption of all single family homes in Cedar Falls is calculated from the recorded electric utility bills in 2010. The measured average monthly and yearly electricity consumption are compared with the expected yearly and monthly electricity consumption obtained from the GA-NMM model. Guidelines on the validation of building energy simulation models is currently based on a model's compliance with standard criteria for Coefficient of Variation of Root Mean Square Error (CVRMSE) (%) and Mean Bias Error (MBE) (%), which are obtained from Equations (3) and (4). The acceptable values for MBE and CVRMSE given in ASHRAE Guideline 14 for monthly time interval are 5% and 15% respectively [62,63].

$$MBE(\%) = \frac{\sum_{i=1}^{N_p}(m_i - s_i)}{\sum_{i=1}^{N_p}(m_i)} \tag{3}$$

$$CV\ RMSE(\%) = \frac{\sqrt{(\sum_{i=1}^{N_p}(m_i - s_i)^2 / N_p)}}{\bar{m}} \tag{4}$$

3.2.1 Obtaining monthly electricity consumption from electricity bills

The original dataset of billing data used in this study includes metered residential dwellings, condominiums/townhouses, two-family dwellings (duplexes), multiple-family dwellings, and apartment units. In this study, since the focus of this study is on single family homes, the multi-family buildings in this dataset, including some of which do not individually meter their electricity use at each unit, are excluded.

The time period that is used for this study is from January to December 2010. Therefore, the bills which have measurements during and after January 1st, 2010 are considered in this analysis. It is also noted that in working with utility billing data, the billing data durations vary from 5 to 90 days. However, since more than 90% of the bill's duration are between 25 to 35 days, the bills outside this range of 25 to 35 days are excluded from the dataset. After applying these adjustments and excluding the outliers, from the original bill records, a dataset of 101,220 electric bill records for 8,370 single family buildings in 2010 are ultimately utilized.

Using the refined dataset, it is also noted that the start and end dates of each billing cycle are not uniform, as is typical for residential building billing in the U.S. To support comparison of electricity consumption across uniform time periods, the electricity consumption recorded for each bill is divided by the billing duration and then multiplied by the number of days in each month. For example, if a bill starts on January 26 and ends on March 2, hence, it covers 5 days of January, 28 days of February and 2 days of March, then the number of days in each month will be multiplied with average daily electricity consumption obtained from bill amount divided by bill duration.

4 Results and Discussion

The GA-NMM technique is used to generate the three PMF positions with their weights for each of the four selected variables. Using the joint distribution function that combines all the PMFs, 9 index buildings were developed with their associated weights. The energy models were developed for each of the defined index buildings and the annual and monthly site electricity consumption of the index buildings and their associated weights were calculated based on the energy model results. Using a weighted summation, the expected annual and monthly site electricity consumption of the single family buildings in Cedar Falls are then estimated and reported in this section, and compared to the measured data.

4.1 GA-NMM model results

Applying the GA-NMM technique for the four main building variables (area, cooling efficiency, attic insulation R-value, window type), results in 3 PMF positions with their associated weights for each variable, as shown in Table 1.

Table 1. PMF positions for each of the four building variables utilized in GA-NMM

Variables	L1	L2	L3	w_1	w_2	w_3
Area (m^2)	95.1	161	264	0.46	0.47	0.08
Cooling Efficiency (SEER)	7	11	15	0.15	0.62	0.23
Attic insulation R-value	13	27	47	0.27	0.52	0.21
Window type	1	2	3	0.30	0.31	0.39

Note: For window type, 1, 2, and 3 represent single pane (clear), double pane (clear, metal frame, air filled), and double pane (low E, nonmetal frame, air filled)

These discrete distribution functions for each variable need to be combined to generate a junction distribution, which consists of nine indices that statistically reflect the building population characteristics. Table 2 shows the characteristics of the final nine index buildings that statistically represent the population. Since the summation of the weights should be equal to one, the weigh for index 1 is calculated using the normality condition,

$w_1 = 1 - \sum_2^9 w_i$. This condition may result in a “negative” weight value for index 1 that is mathematically derived to match the desired moments of original data, however, it does not have a physical correspondence.

The energy model for each of these index buildings was then developed separately using EnergyPlus. The annual site electricity consumption for each index building and their associated weights are given in Table 3.

Table 2. Index building characteristics

Index	Area (m^2)	Cooling Efficiency (SEER)	Attic insulation R-value	Window type
1	161.1	11	27	2
2	95.1	11	27	2
3	263.6	11	27	2
4	161.1	7	27	2
5	161.1	15	27	2
6	161.1	11	13	2
7	161.1	11	47	2
8	161.1	11	27	1
9	161.1	11	27	3

Note: For window type, 1, 2, and 3 represent single pane (clear), double pane (clear, metal frame, air filled), and double pane (low E, nonmetal frame, air filled)

Table 3. Annual site electricity consumption of the index buildings with their associated weights

Index	Weights	Site Electricity consumption kWh/yr
1	-1.079	11,040
2	0.456	9,601
3	0.076	12,110
4	0.150	11,527
5	0.227	10,281
6	0.266	11,169
7	0.211	10,832
8	0.300	11,158

9	0.393	10,598

Using Equations (1) and (2), the expectation of yearly site electricity consumption of single family buildings in Cedar Falls is calculated to be 10,219 kWh, with standard deviation of 786 kWh per year. The same procedure are applied to calculate the expected monthly electricity consumption and their associated standard deviation which are reported in Table 4.

Table 4. Expected monthly site electricity consumption for a single family buildings in Cedar Falls and their standard deviation

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Expected monthly electricity consumption (kWh)	979	859	776	687	791	849	1137	1180	769	702	737	920
Standard Deviation (kWh)	119	102	85	66	68	85	153	175	60	66	80	108

4.2 GA-NMM Model Validation

The average monthly measured electricity consumption for a single family building in Cedar Falls is compared with the expected electricity consumption from the model results; the percentage difference for each month is reported in Table 5. MBE and CV-RMSE are calculated as 4.5% and 7.8%, respectively, which is below the recommended maximum monthly acceptance criteria in ASHRAE Guideline 14 [63]. This indicates that there is reasonable agreement between measured data and GA-NMM model results. The proposed methodology is applied for the year 2010 and validated with the electricity consumption of the single family homes in Cedar Falls for this year. The performance of the model in other years has not been investigated in this research due to lack of a complete dataset of electricity consumption measured data in other years. Further investigation of the performance of this model in other years is the subject of future work.

Reviewing other building modeling studies also indicate that the performance of GA-NMM model is acceptable, and performs similar or better. For example, in a study by Harberl et al [64], a four zone, single story electrically heated and cooled building was simulated and compared with hourly measured electricity. MBE of -0.7% and CV-RMSE of 23.1% are reported for their model results. In another study by Wan and Yik [65], a model representing typical residential buildings in Hong Kong is developed to predict the annual and monthly electricity consumption and compared with data found from surveys of building characteristics and relevant public statistics. Based on the reported results, the estimated overall annual electricity consumption is higher than measured data by 44%.

Table 5. Monthly electricty use comparision between GA-NMM model results and measured data

Monthly electricity consumption (kWh)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Measured data	967	813	770	665	798	1001	1263	1230	847	735	810	973
GA-NMM model results	979	859	776	687	791	849	1138	1180	769	702	738	920
Difference %	-1%	-6%	-1%	-3%	1%	15%	10%	4%	9%	5%	9%	6%

With the expected standard deviation from the expected electricity consumption for each month, the uncertainty range for the model results is calculated by adding and subtracting the calculated STD, then compared with the measured data. As shown in Figure 2, in most cases the measured data is within the expected band for electricity consumption obtained from the model. It is noted that the model under-predicts the electricity use in June and September in comparison to other months. This may be due to a slight under prediction in the non-weather dependent loads during these periods.

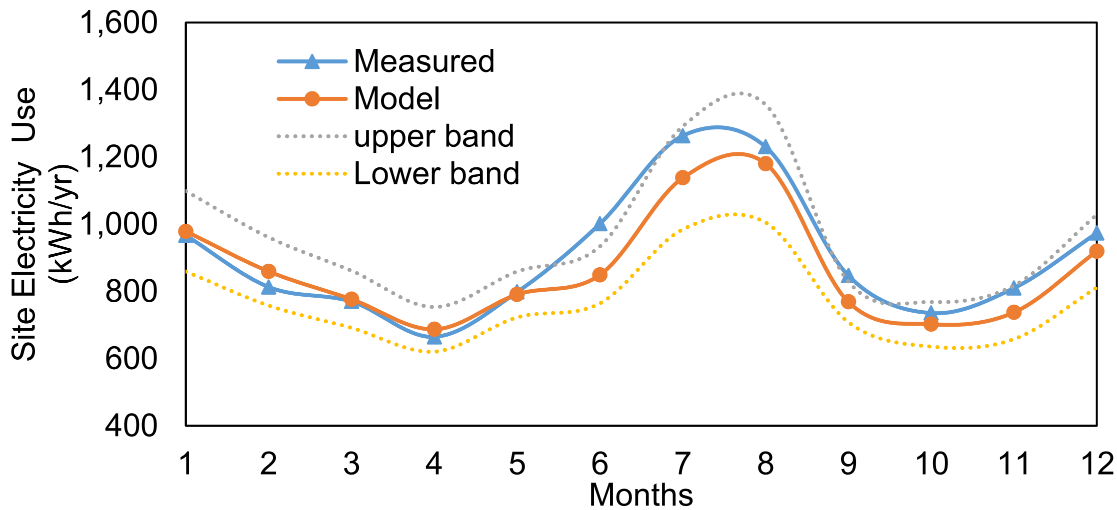

Figure 2. Measured monthly electricity consumption data compared to the expected site electricity consumption and the confidence interval band obtained from model. The upper band is $E[x] + STD$ and the lower band is $E[x] - STD$.

5 Conclusions

In this study, a new technique is proposed to predict the annual and monthly site electricity consumption for single family homes by utilizing the GA-NMM method, a sampling method that develops a small group of index buildings and associated weighting criteria. These index buildings statistically represent the energy-related features of the larger dataset of buildings. Building energy models of the index buildings were developed to predict the annual and monthly site electricity consumption of the building population. This methodology was applied to predict the annual and monthly electricity consumption use of 8,370 single family homes in Cedar Falls, IA. The following main conclusions can be drawn from this study:

- Based on data available from energy audit information and assessor’s data, four influential variables, including area, cooling system efficiency, attic insulation R-value and window type were found to be available and influential variables for use as inputs into the model development.
- With out-of-sample data, which includes electricity use data for all homes except those used to develop the model index buildings, the annual electricity consumption predicted is 10,219 kWh, which is within 6% of the measured electricity consumption for single family buildings in Cedar Falls. The predicted monthly electricity consumption MBE and CV-RMSE are 4.5% and 7.8% respectively, which are within the acceptable range based on ASHRAE Guideline14. Thus for these frequencies of data, this method has acceptable performance, with relatively low computational effort.

Moving forward, with additional data, including smart meter data, this method can be further tested to evaluate performance at a broader range of data frequencies. In addition, particularly for applications where limited data is available about housing energy-related characteristics across a broad set of homes, evaluating the level of accuracy of performance which can be achieved with the use of various input variables may also be beneficial. The proposed method is capable of predicting a reasonably accurate estimation of the energy behavior of the studied population by utilizing the characteristics of local buildings. The main advantage of the proposed method is preserving the accuracy of the results while the computational time and cost is decreased. Furthermore, other outcomes from the energy model such as site energy consumption and source energy consumption, including those with monthly and yearly time resolution, can be estimated for the studied population using the proposed method. As proposed, this effort demonstrates the use of a low-computational intensity method for evaluating energy performance of a set of residential buildings, which can be extrapolated to the city-level and be used by decision makers, utilities and other stakeholders interested in assessing the energy performance of an urban area.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant # 1662691. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] C.F. Reinhart, C. Cerezo Davila, Urban building energy modeling - A review of a nascent field, Build. Environ. 97 (2016) 196–202. doi:10.1016/j.buildenv.2015.12.001.

[2] X. Lü, T. Lu, C.J. Kibert, M. Viljanen, Modeling and forecasting energy consumption for heterogeneous buildings using a physical-statistical approach, Appl. Energy. 144 (2015) 261–275. doi:10.1016/j.apenergy.2014.12.019.

[3] F. Ascione, N. Bianco, R.F. De Masi, F. de' Rossi, G.P. Vanoli, Energy refurbishment of existing buildings through the use of phase change materials: Energy savings and indoor comfort in the cooling season, Appl. Energy. 113 (2014) 990–1007. doi:10.1016/j.apenergy.2013.08.045.

[4] H. Do, K.S. Cetin, Evaluation of the causes and impact of outliers on residential

building energy use prediction using inverse modeling, Build. Environ. 138 (2018) 194–206. doi:10.1016/j.buildenv.2018.04.039.

[5] H. Do, K.S. Cetin, Residential Building Energy Consumption : a Review of Energy Data Availability , Characteristics , and Energy Performance Prediction Methods, Curr. Sustain. Energy Reports. (2018) 76–85.

[6] A. Tsanas, A. Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, Energy Build. 49 (2012) 560–567. doi:10.1016/j.enbuild.2012.03.003.

[7] D. Mitra, Y. Chu, N. Steinmetz, K.S. Cetin, P. Kremer, J. Lovejoy, Defining Typical Occupancy Schedules and Behaviors in Residential Buildings Using the American Time Use Survey, in: ASHRAE Annu. Conf. Kansas City, MO, June 22-26, 2019., 2019.

[8] J.A. Fonseca, A. Schlueter, Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts, Appl. Energy. 142 (2015) 247–265. doi:10.1016/j.apenergy.2014.12.068.

[9] S. Heiple, D.J. Sailor, Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles, Energy Build. 40 (2008) 1426–1436. doi:10.1016/j.enbuild.2008.01.005.

[10] W. Li, Y. Zhou, K.S. Cetin, S. Yu, Y. Wang, B. Liang, Developing a landscape of urban building energy use with improved spatiotemporal representations in a cool-humid climate, Build. Environ. 136 (2018) 107–117. doi:10.1016/j.buildenv.2018.03.036.

[11] R. Nouvel, A. Mastrucci, U. Leopold, O. Baume, V. Coors, U. Eicker, Combining GIS-based statistical and engineering urban heat consumption models: Towards a new framework for multi-scale policy support, Energy Build. 107 (2015) 204–212. doi:10.1016/j.enbuild.2015.08.021.

[12] D.J. Sailor, L. Lu, A top-down methodology for developing diurnal and seasonal anthropogenic heating profiles for urban areas, Atmos. Environ. 38 (2004) 2737–2748. doi:10.1016/j.atmosenv.2004.01.034.

[13] J. Huang, E. Franconi, Commercial heating and cooling loads component analysis, Environ. Energy Technol. Div. Lawrence Berkeley Natl. Lab. Berkeley, CA. (1999) 1–133.

[14] L.G. Swan, V.I. Ugursal, Modeling of end-use energy consumption in the residential sector: A review of modeling techniques, Renew. Sustain. Energy Rev. 13 (2009) 1819–1835. doi:10.1016/j.rser.2008.09.033.

[15] M. Kavgic, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, M. Djurovic-Petrovic, A review of bottom-up building stock models for energy consumption in the residential sector, Build. Environ. 45 (2010) 1683–1697. doi:10.1016/j.buildenv.2010.01.021.

[16] R. Ghedamsi, N. Settou, A. Gouareh, A. Khamouli, N. Saifi, B. Recioui, B. Dokkar, Modeling and forecasting energy consumption for residential buildings in Algeria using bottom-up approach, Energy Build. 121 (2016) 309–317. doi:10.1016/j.enbuild.2015.12.030.

[17] K. Han, J.S. Haberl, Development of methodology for calibrated simulation in

single-family residential buildings using three-parameter change-point regression model, Energy Build. 99 (2015) 140–152. doi:10.1016/j.enbuild.2015.04.032.

[18] K.X. Perez, K. Cetin, M. Baldea, T.F. Edgar, Development and analysis of residential change-point models from smart meter data, Energy Build. 139 (2017) 351–359.

[19] M.S. Ahmed, A. Mohamed, R.Z. Homod, H. Shareef, Hybrid LSA-ANN Based Home Energy Management Scheduling Controller for Residential Demand Response Strategy, Energies. (2016). doi:10.3390/en9090716.

[20] Y. Zhang, Z.O. Neill, B. Dong, G. Augenbroe, Comparisons of inverse modeling approaches for predicting building energy performance, Build. Environ. 86 (2015) 177–190. doi:10.1016/j.buildenv.2014.12.023.

[21] M. Castelli, L. Trujillo, L. Vanneschi, Prediction of energy performance of residential buildings : A genetic programming approach, Energy Build. 102 (2015) 67–74. doi:10.1016/j.enbuild.2015.05.013.

[22] Z.O. Neill, C.O. Neill, Development of a probabilistic graphical model for predicting building energy performance, Appl. Energy. 164 (2016) 650–658. doi:10.1016/j.apenergy.2015.12.015.

[23] R. Li, Z. Wang, C. Gu, F. Li, H. Wu, A novel time-of-use tariff design based on Gaussian Mixture Model q, Appl. Energy. 162 (2016) 1530–1536. doi:10.1016/j.apenergy.2015.02.063.

[24] R.K. Jain, K.M. Smith, P.J. Culligan, J.E. Taylor, Forecasting energy consumption

of multi-family residential buildings using support vector regression : Investigating the impact of temporal and spatial monitoring granularity on performance accuracy, Appl. Energy. 123 (2014) 168–178. doi:10.1016/j.apenergy.2014.02.057.

[25] M. Aydinalp, V.I. Ugursal, A.S. Fung, Modeling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks, Appl. Energy. 79 (2004) 159–178. doi:10.1016/j.apenergy.2003.12.006.

[26] M. Aydinalp, V. Ismet Ugursal, A.S. Fung, Modeling of the appliance, lighting, and space-cooling energy consumptions in the residential sector using neural networks, Appl. Energy. (2002). doi:10.1016/S0306-2619(01)00049-6.

[27] M. Parti, The Total and Appliance-Specific Conditional Demand for Electricity in the Household Sector Author (s): Michael Parti and Cynthia Parti Published by : RAND Corporation Stable URL : http://www.jstor.org/stable/3003415 The total and appliance-specific con, 11 (1980) 309–321.

[28] D.J. Aigner, C. Sorooshian, P. Kerwin, S. The, E. Journal, N. July, Conditional Demand Analysis for Estimating Residential End-Use Load Profiles Published by : International Association for Energy Economics Conditional Demand Analysis for Estimating Residential End-Use Load Profiles, 5 (2016) 81–97.

[29] E. Hirst, R. Goeltz, D. White, Determination of household energy using 'fingerprints' from energy billing data, Int. J. Energy Res. 10 (1986) 393–405.

[30] G. Raffio, O. Isambert, G. Mertz, C. Schreier, K. Kissock, Targeting residential energy assistance, in: ASME 2007 Energy Sustain. Conf., 2009: pp. 489–495.

[31] M. Kavgic, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, M. Djurovic-petrovic, A review of bottom-up building stock models for energy consumption in the residential sector, Build. Environ. 45 (2020) 1683–1697. doi:10.1016/j.buildenv.2010.01.021.

[32] M.A. McNeil, V.E. Letschert, Modeling diffusion of electrical appliances in the residential sector, Energy Build. (2010). doi:10.1016/j.enbuild.2009.11.015.

[33] A. Capasso, W. Grattieri, R. Lamedica, A. Prudenzi, A bottom-up approach to residential load modeling, IEEE Trans. Power Syst. 9 (1994) 957–964. doi:10.1109/59.317650.

[34] A. Parekh, Development of archetypes of building characteristcs libraries for simplified energy use evaluation of houses, Proc. BS2005 Ninth Int. IBPSA Conf. (2005) 921–928. doi:10.1.1.470.363.

[35] W.A. MacGregor, F. Hamdullahpur, V.I. Ugursal, Space heating using small-scale fluidized beds: A technoeconomic evaluation, Int. J. Energy Res. 17 (1993) 445–466.

[36] H. Farahbakhsh, V.I. Ugursal, A.S. Fung, A residential end-use energy consumption model for Canada, Int. J. Energy Res. 22 (1998) 1133–1143.

[37] R.L. Wasserstein, Monte carlo: Concepts, algorithms, and applications, (1997).

[38] R.E. Caflisch, Monte carlo and quasi-monte carlo methods, Acta Numer. 7 (1998) 1–49.

[39] I.H. Cho, K. Porter, Modeling building classes using moment matching, Earthq.

Spectra. 32 (2016) 285–301. doi:10.1193/071712EQS239M.

[40] I. Iervolino, G. Manfredi, M. Polese, G.M. Verderame, G. Fabbrocino, Seismic risk of R.C. building classes, Eng. Struct. 29 (2007) 813–820. doi:10.1016/j.engstruct.2006.06.019.

[41] U.S. Energy Information Administration, U.S. electricity flow, (2017). https://www.eia.gov/totalenergy/data/monthly/pdf/flow/electricity.pdf (accessed January 3, 2020).

[42] U.S. Energy Information Administration, Monthly Energy Review, 2019. doi:10.1044/leader.ppl.24062019.28.

[43] U.S. Energy Information Administration, Residential Energy Consumption Survey(RECS), (2013). https://www.eia.gov/consumption/residential/data/2009/index.php?view=consumption#undefined.

[44] U.S. Energy Information Administration, Commercial Buildings Energy Consumption Survey (CBECS), (2018). https://www.eia.gov/consumption/commercial/ (accessed January 3, 2020).

[45] National Renewable Energy Laboratory (NREL), ResStock, US Dep. Energy. (2018). https://www.nrel.gov/buildings/resstock.html (accessed January 3, 2020).

[46] National Renewable Energy Laboratory (NREL), ComStock, (2019). https://www.nrel.gov/buildings/commercial.html (accessed July 20, 2019).

[47] National Renewable Energy Laboratory (NREL), URBANopt Advanced Analytics

Platform, (2018). https://www.nrel.gov/buildings/urbanopt.html (accessed January 4, 2020).

[48] C. Cerezo, T. Dogan, J.A. Jakubiec, T. Rakha, C. Rose, Urban Modeling Interface (UMI), 2018. (n.d.). http://web.mit.edu/sustainabledesignlab/projects/umi/index.html (accessed November 20, 2019).

[49] A. Energy, ECAD Ordinance, (2018). https://austinenergy.com/ae/energy-efficiency/ecad-ordinance/ecad-for-residential-customers (accessed April 20, 2018).

[50] I.H. Cho, I. Song, Y.L. Teng, Numerical moment matching stabilized by a genetic algorithm for engineering data squashing and fast uncertainty quantification, Comput. Struct. 204 (2018) 31–47. doi:10.1016/j.compstruc.2018.04.002.

[51] United States Census Bureau, American Housing Survey, (2019). https://www.census.gov/programs-surveys/ahs.html (accessed September 27, 2019).

[52] E. Jahani, V. Soham, J. David, G. William, C. Kristen, City-Scale High-Resolution WRF-UCM Urban Weather Predictions Compared to a Dense Network of Ground-Based Weather Station Data for Assessment of Urban Building Energy Consumption, in: ASHRAE Annu. Conf. Kansas City , MO, ASHRAE, Kansas City, MO, USA, 2019.

[53] White Box Technology, Weather Data For Energy Calculation, (2018). http://weather.whiteboxtechnologies.com/hist (accessed January 3, 2020).

[54] R. and A.-C.E. American Society of Heating, Standard 90.1-2013. Energy Standard for Buildings Except Low-Rise Residential Buildings, Am. Soc. Heating, Refrig. Air-Conditioning Eng. Inc. 2013 (2013) 278. doi:http://dx.doi.org/10.1108/17506200710779521.

[55] J. Ching, K.A. Porter, J.L. Beck, Propagating uncertainties for loss estimation in performance-based earthquake engineering using moment matching, Struct. Infrastruct. Eng. 5 (2009) 245–262. doi:10.1080/15732470701298323.

[56] C.L. Karr, B. Weck, L.M. Freeman, Solutions to systems of nonlinear equations via a genetic algorithm, Eng. Appl. Artif. Intell. 11 (1998) 369–375. doi:10.1016/S0952-1976(97)00067-5.

[57] U.S. Energy Information Administration, Annual household site end-use consumption by fuel in the U.S.—totals, 2015, 2018. https://www.eia.gov/consumption/residential/data/2015/c&e/pdf/ce4.1.pdf.

[58] U.S. Department of Energy, EnergyPlus, (n.d.).

[59] National Renewable Energy Laboratory (NREL), BEopt, (2019). https://beopt.nrel.gov/ (accessed April 19, 2018).

[60] E. Wilson, C.. Engebrecht Metzger, S. Horowitz, R. Hendron, 2014 Building America House Simulation Protocols (NREL/TP-5500-60988), (2014).

[61] R. Hendron, C. Engebrecht, Building America House Simulation Protocols, (2010) 79. doi:10.1017/CBO9781107415324.004.

[62] D. Coakley, P. Raftery, M. Keane, A review of methods to match building energy

simulation models to measured data, Renew. Sustain. Energy Rev. 37 (2014) 123–141. doi:10.1016/j.rser.2014.05.007.

[63] American Society of Heating; Refrigerating and Air-Conditioning Engineers, Guideline 14-2002, 2002.

[64] T.E.B.-S.& J.S. Haberl, AN IMPROVED PROCEDURE FOR DEVELOPING CALIBRATED HOURLY SIMULATION MODELS Tarek E . Bou-Saada & Jeff S . Haberl Energy Systems Laboratory Texas A & M University System College Station , TX Photocpied with permission from the International Building Performa, (1995).

[65] K.S.Y. Wan, F.H.W. Yik, Representative building design and internal load patterns for modelling energy use in residential buildings in Hong Kong, 77 (2004) 69–85. doi:10.1016/S0306-2619(03)00104-1.

Appendix

In this section the distribution of each variable utilized in this work is provided. Four variables which are (area, cooling efficiency, attic insulation R-value, window type) investigated in this study are distributed as shown in Figure A1.

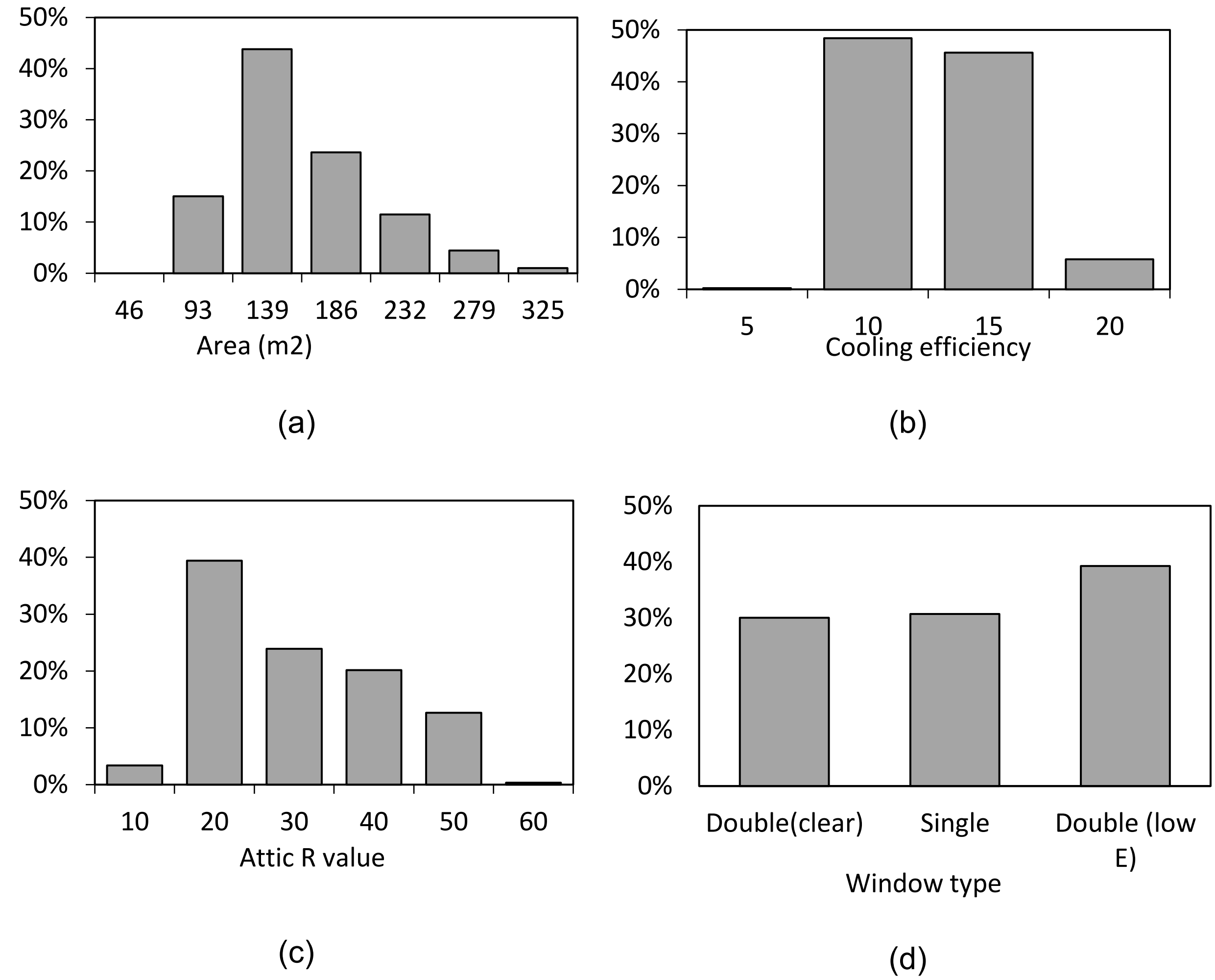

Figure A1. Characteristics of variables : a) building area (m^2), b) cooling system efficiency, c) attic insulation R-value, d) window type

The input values to the GA-NMM model, which include the maximum and minimum of each variable (Table A1) and the calculated first five moments, including area, cooling efficiency, attic insulation R-value (Table A2), are also presented. It should be noted that for window type since the original data is categorical (single pane (clear), double pane (clear, metal frame, air-filled), and double pane (low-E, non-metal frame, air-filled) its original distribution is used as the probability mass function (PMF) positions.

Table A1 Maximum and minimum values of the variables

Variables	Min	Max
Area (m^2)	49	312
Cooling Efficiency (SEER)	2	19
Attic insulation R-value	0	58

Table A2 First five moments of the variables

Variables	$E(x)$	$E(x^2)$	$E(x^3)$	$E(x^4)$	$E(x^5)$
Area (m^2)	1.49E+03	2.50E+06	4.67E+09	9.67E+12	2.19E+16
Cooling Efficiency (SEER)	1.11E+01	1.30E+02	1.59E+03	2.01E+04	2.64E+05
Attic insulation R-value	2.76E+01	8.99E+02	3.30E+04	1.32E+06	5.60E+07