Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 143 (2017) 859–864

Energy Procedia

www.elsevier.com/locate/procedia

World Engineers Summit – Applied Energy Symposium & Forum: Low Carbon Cities & Urban Energy Joint Conference, WES-CUE 2017, 19–21 July 2017, Singapore

A Generalizable Method for Estimating Household Energy by Neighborhoods in US Urban Regions

Wenwen Zhang[a], Subhrajit Guhathakurta[a]*, Ram Pendyala[b], Venu Garikapati[b], and Catherine Ross[c]

[a] *Center for Geographic Information Systems, Georgia Tech, 760 Spring St. NW, Atlanta, GA 30308, USA*
[b] *School of Sustainable Engineering and the Built Environment, Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, AZ 85287-3005, USA*
[b] *Center for Quality Growth and Regional Development, Georgia Tech, 760 Spring St. NW, Atlanta, GA 30308, USA*

Abstract

There is mounting evidence to suggest that the urban built form plays a crucial role in household energy consumption, hence planning energy efficient cities requires thoughtful design at multiple scales - from buildings, to neighborhoods, to urban regions. While data on household energy use are essential for examining the energy implications of different built forms, few utilities providing power and gas offer such information at a granular scale. Therefore, researchers have used various estimation techniques to determine household and neighborhood scale energy use. In this study we develop a novel method for estimating household energy demand that can be applied to any urban region in the US with the help of publicly available data. To improve estimates of residential energy this paper describes a methodology that utilizes a matching algorithm to stitch together data from RECS with the Public Use Microdata Sample (PUMS) provided by the Bureau of Census. Our workflow statistically matches households in RECS and PUMS datasets based on the shared variables in both, so that total energy consumption in the RECS dataset can be mapped to the PUMS dataset. Following this mapping procedure, we generate synthetic households using processed PUMS data together with marginal totals from the American Community Survey (ACS) records. By aggregating energy consumptions of synthesized households, small area or neighborhood-based estimates of residential energy use can be obtained.

Peer-review under responsibility of the scientific committee of the World Engineers Summit – Applied Energy Symposium & Forum: Low Carbon Cities & Urban Energy Joint Conference.

Keywords: energy modeling; residential energy; machine learning; household synthesis; statistical matching

* Corresponding author. Tel.: +1 (404) 385-0900; fax: +1 (404) 385-0450.
E-mail address: subhro@gatech.edu

1876-6102
Peer-review under responsibility of the scientific committee of the World Engineers Summit – Applied Energy Symposium & Forum: Low Carbon Cities & Urban Energy Joint Conference.
10.1016/j.egypro.2017.12.774

1. Introduction

The increasing concerns around fossil fuel use and its climate-changing GHG emissions have generated an impetus for planning and retrofitting neighborhoods to be more energy efficient. Reducing residential energy use is a significant component of the global response to climate change given that the residential sector consumed about 21 percent of total energy in the U.S. (USEIA 2017). However, estimating residential energy consumption is challenging especially because the utility providers are often unwilling to share this information at a granular spatial scale. As a result, estimates of residential energy demand is based on several modelling techniques, with various margins of error. These models are influenced by several factors, including characteristics of housing units, occupant behavior, and adjacent built environment attributes. Typically, household surveys are used to obtain information regarding these factors. The surveys additionally often request actual energy use data from householders' utility bills. While the surveys offer a viable option for generating estimates of residential energy consumption, they are time-consuming and expensive. Besides, the United States Energy Information Administration (USEIA) conduct detailed surveys of energy use in the residential sector every four years, which also includes a wealth of information about many of the factors mentioned earlier influencing energy use.

In this paper we propose a new bottom-up residential operational energy consumption model using the USEIA's Residential Energy Consumption Survey (RECS), the Public Use Microdata Sample (PUMS), and the American Community Survey (ACS) data as inputs. These data are available for all the major metropolitan areas in the U.S. Therefore, the methodology presented in this paper can potentially be applied to estimate residential energy consumption for all major cities. The remainder of the paper is organized as follows. The next section outlines some prior work in estimating residential energy using top-down and bottom-up models. Section Three describes our proposed model framework, followed by an implementation of the proposed model using Atlanta Metropolitan Area in Section Four. The model results are presented in the fifth section and validated using utility data for the City of Atlanta. The last part concludes the proposed residential operation energy consumption model and identifies constraints which merit further examination and improvements.

2. Prior work

Studies estimating operational energy consumption primarily use either top-down or bottom-up models. The top-down models typically estimate local residential energy consumption from regional level estimates using factors, such as gross domestic product (Eric Hirst, 1978; Saha & Stephenson, 1980), technology attributes (Haas & Schipper, 1998; E. Hirst, Lin, & Cope, 1977), price, total population, and evolution of housing stocks (Nesbakken, 1999; Zhang, 2004). The bottom-up approach consists two genres of models, including the statistical models and the engineering models. The results from statistical models are generally analyzed to interpret the correlations of energy consumption with various individual-level characteristics, such as the size of housing units, socio-economic and demographic features, local heating and cooling degree days (Eric Hirst, Goeltz, & White, 1986; Raffio, Isambert, Mertz, Schreier, & Kissock, 2007) and behaviors, including financial and cultural motivations in energy use (Douthitt, 1989; Fung, Aydinalp, & Ugursal, 1999; Tonn & White, 1988). The aim of statistical models is to understand the variation in energy use given changes in various occupant characteristics, so that policies can be derived to monitor energy price and provide ethical or financial motivations to regulate or curb energy consumptions (Chen, Wang, & Steemers, 2013; Jain, Smith, Culligan, & Taylor, 2014). Some recent model efforts use more advanced machine learning models such as neural networks and support vector machine to estimate residential energy consumptions (Aydinalp, Ugursal, & Fung, 2002, 2004; Dong, Cao, & Lee, 2005). However, these models are also data intensive and are difficult to be applied to large study areas.

The engineering models compute energy consumptions based on the energy ratings of various appliances, building materials, applied energy saving technology on-site and thermodynamic theorems (Zhao & Magoulès, 2012). Specifically, this approach first estimates energy consumption for a series of typical prototypes or archetypes of housing stocks in the region, using a small sample of buildings. The occupant behaviors are not captured but simplified to various assumptions. Different from the statistical models, the objective of engineering models is to extrapolate the results to the entire region so that the total residential energy consumption or the changes in consumption under various technology penetration scenarios can be obtained. The extrapolation is usually made by assigning weights, estimated based on regional housing inventory, to the sampled buildings. several software and calculation standards are

developed to estimate building energy consumption using this modeling approach (Crawley, Hand, Kummert, & Griffith, 2008; de Normalización, 2008)

Despite different model objectives, both the statistical and engineering models have some limitations. First, both approaches rely heavily on the availability of historical micro sample data, which are usually time-consuming and expensive to collect (Swan & Ugursal, 2009; Zhao & Magoulès, 2012). For instance, the U.S. EIA conducts national level Residential Energy Consumption Survey (RECS) every four years to collect information regarding occupant characteristics, market penetration of new appliances and condition of housing stocks. The sample size, however, in each metropolitan area is quite small, rendering it insufficient to assess local energy consumptions. Additionally, it is hard, if not impossible, to extrapolate the micro-sample model results to the region, as the features considered in the models are usually not standard across the region (Kavgic et al., 2010).

3. Estimation Technique

The residential operational energy estimation technique developed in this study is based on three key components: 1) statistical matching of the household records from the RECS and PUMS data, 2) estimation of energy consumption models using the matched records, and 3) synthesizing households using PUMS and ACS data. The objective of the first component is to assign the energy use component available in RECS to a matched record in the PUMS dataset. In other words, the statistical matching process "joins" energy consumption variable from the RECS data to a portion of the PUMS data based on the similarity of the shared variables. Given that RECS has a substantially smaller sample than PUMS, a lot of records in PUMS remain unmatched. The second model component then estimates models using energy consumption variables from the RECS data as the dependent variables and household socio-demographic and economic features and housing unit characteristics in the PUMS data as the independent variables. The estimated model is then applied to the unmatched PUMS data to impute energy consumptions. The final output of the second model component is the PUMS data with appended energy consumption variables.

The final model component takes energy consumption appended to PUMS data as the seed matrix and ACS data as the marginal controls to synthesize households in the region. Both household-level and population-level variables, which are highly correlated with energy consumptions in the estimated models, are controlled to obtain a more representative profile for the region. The synthesized energy consumptions can then be aggregated using various geographic units to determine the distribution of energy consumptions in the region. The following sections elaborate the detailed methodologies regarding each model component.

4. Data and Model Implementation

We implemented the estimation technique described above using data for 10-county Atlanta metropolitan region. The results are then cross-compared with electricity and natural gas consumption data provided by Georgia Power and Atlanta Gas Light to validate the model outputs.

4.1. Statistical Matching

The RECS data and PUMS data are matched for all records in the state of Georgia using common variables in both datasets, so that the energy consumption information, including electricity BTU (ELBTU), natural gas BTU (NGBTU) and other BTU (OBTU), can be appended to the matched PUMS records. First, we identified 12 variables, including the type of housing unit, property ownership, year built, energy bills, etc., from the RECS data that are also available in the PUMS data. These variables, however, are not exactly measuring the same attribute in the same manner as evident from the data dictionaries. To unify the measurements, the common variables were reclassified or reorganized. We conducted the Spearman rho rank correlation analysis to determine the final matching variables, and the results are shown in Table 1. The variables with the highest correlations (bolded) are used as matching variables. The household structure and heating fuel are categorical variables and, therefore, are used as group control variables, indicating that only households with the same household structure and the heating fuel type can be joined. The joint distribution of household type and heating fuel type reveals that some heating fuel types, such as solar, district steam, and coal, are only used by a small sample of households. As a result, there will be many housing categories with zero observations, if all heating fuel types are persevered in the model, rendering the matching process infeasible to complete. For instance, if there is no apartment with wood as heating fuel in the RECS data, then all similar records from the PUMS will not be able to be matched and appended with energy

consumptions. Therefore, to simplify the matching process, the heating fuel types are reclassified into three categories: 1) electricity, 2) natural gas, and 3) others.

The continuous variables, such as the total number of rooms, annual electricity, natural gas, and other bills are used as match variables. To eliminate the influence of various measuring units on the Manhattan Distance calculation outcome, all the continuous matching variables are standardized. Different matching variables are used to join ELBTU, NGBTU, and OBTU to the PUMS data, as shown in Table 2. The number of bedrooms is not used as a matching variable to joining ELBTU, as it is highly correlated with the total number of rooms in both datasets.

Table 1: Spearman rho rank correlation analysis results

	Variable names	$Adjusted\ \rho^2[p\ value]$			N
		Electricity BTU	Natural Gas BTU	Other BTU	
Categorical	Household Structure	**0.232 [0.00]**	0.034 [0.00]	0.037 [0.00]	2246
	Tenure Type	0.095 [0.00]	0.010 [0.00]	0.022 [0.00]	2246
	Heat Fuel Type	0.046 [0.00]	**0.515 [0.00]**	**0.179 [0.00]**	2246
	Income	0.110 [0.00]	0.040 [0.00]	0.000 [0.40]	2246
	Move in time	0.030 [0.00]	-0.001 [0.58]	0.033 [0.00]	2246
	Year Built	0.013 [0.00]	0.036 [0.00]	0.016 [0.00]	2246
Numerical	Bedrooms	0.261 [0.00]	0.073 [0.00]	0.005 [0.00]	2215
	Total Rooms	**0.249 [0.00]**	0.079 [0.00]	0.009 [0.00]	2246
	Household Size	0.212 [0.00]	0.009 [0.00]	0.001 [0.21]	2246
	Annual Electric Bill	**0.894 [0.00]**	0.027 [0.00]	0.003 [0.02]	2246
	Annual Natural Gas Bill	0.037 [0.00]	**0.996 [0.00]**	0.028 [0.00]	2246
	Annual Other Bill	0.006 [0.00]	0.027 [0.00]	**0.999 [0.00]**	2246

Table 2: Sets of Matching Variables by Target Variable

Type of Matching Variables (X_M)	Target Variables (Z)		
	Electricity BTU	**Natural Gas BTU**	**Other BTU**
Group Control Variables	Household Structure	Heat Fuel Type	Heat Fuel Type
Distance Calculation Variables	Total Rooms Annual Electricity Bill	Annual Natural Gas Bill	Annual Other Bill

The records from RECS and PUMS are then matched together by minimizing the differences between distance calculation variables for households with the same category in the group control variables. There are significantly more samples in the PUMS data, compared with the RECS data. Therefore, for each RECS dataset, the closest PUMS record is found and matched. For ELBTU, the median distance is 0.038, and 75% percentile distance value is 0.057, indicating the data are well matched together. The maximum distance for matching results of OBTU is 0.98 when two unusual records are eliminated. For NGBTU matching results, the maximum distance is 0.06, suggesting all the households are successfully matched.

4.2. BTU Imputation Model Results

In this step, various machine learning models are estimated using the matched PUMS datasets, and the estimated models are then applied to unmatched PUMS records to impute residential energy consumptions by BTU types. Before model estimation, we first preprocessed features in the PUMS data sets. Features with more than 10% missing values were not considered in the models. This excludes 28 features in the PUMS dataset. Among the remaining features, nine continuous variables were standardized. The rest 36 categorical features were converted into 134 binary features. The number of features included in the model is 143. The averaged results of the 10-fold cross validation experiments for electricity consumption are shown in Table 3. All the training models use the default parameter settings in the Scikit Learn package. The results are sorted by Mean Absolute Errors.

Table 3: Model Results for Electricity Energy Consumption

Models	Mean Absolute Error	Median Absolute Error	R^2	Mean Average Percent Diff.
ElasticNet	**4.34e+03 +/- 164.10**	**6.40e+03 +/- 138.02**	**0.88 +/- 0.01**	**13.38 +/- 0.43**
Lasso	4.70e+03 +/- 267.22	6.68e+03 +/- 165.09	0.87 +/- 0.01	14.09 +/- 0.50
Ridge	4.71e+03 +/- 281.09	6.67e+03 +/- 162.89	0.87 +/- 0.01	14.07 +/- 0.49
Linear	4.73e+03 +/- 291.61	6.71e+03 +/- 168.62	0.87 +/- 0.01	14.20 +/- 0.52
Bagging	4.89e+03 +/- 171.73	7.08e+03 +/- 135.53	0.85 +/- 0.01	14.50 +/- 0.33
Random Forest	4.93e+03 +/- 203.79	6.94e+03 +/- 168.49	0.86 +/- 0.01	14.86 +/- 0.50
Gradient Boosting	5.02e+03 +/- 239.00	6.95e+03 +/- 118.06	0.85 +/- 0.01	14.72 +/- 0.37
AdaBoost	7.19e+03 +/- 303.40	8.62e+03 +/- 205.73	0.82 +/- 0.02	17.06 +/- 0.61
Extra Trees	7.31e+03 +/- 949.76	9.04e+03 +/- 793.19	0.78 +/- 0.03	19.15 +/- 1.53

The outputs for electricity energy consumption suggests that Elastic Net regression performs the best among all tested models, as shown in Table 4. The Elastic Net model outputs present the smallest mean and median absolute errors and average percent difference. The results suggest, on average, the predicted consumptions are approximately 13.4% different from the statistically matched electric BTU consumption values. Additionally, the model also has the largest average R^2 among all the examined models. The results also suggest that other linear models, such as Lasso, Ridge, Ordinary Least Square also show similar prediction power (in terms of the magnitude of errors and R^2), indicating the relationship between explanatory features and target feature (ELBTU) is linear. While, ensemble learning models, such as Bagging, Random Forest, Gradient Boosting, AdaBoost, and Extra Trees, perform comparatively poorly, with much higher absolute errors and lower R^2.

4.3. Household Synthesis

The population synthesizing process is implemented in PopGen 1.1. The software takes PUMS data as the seed matrix, containing joint distributions among various features of households and population. PopGen uses ACS data as the marginal controls to interpolate the weight of each household in the PUMS data. Both household (or housing unit) level and population level variables can be controlled with the IPU synthesizing method. The control variables are selected based on the correlations with the target variables (i.e. ELBTU, NGBTU, and OBTU) and the availability of the data in the ACS data.

The correlation of various features with the target feature is determined using the estimated coefficients from the Elastic Network models and the feature importance scores from the random forest model. The results show the annual electricity and natural gas bills are highly correlated with the electricity and natural gas BTU consumptions, with significantly higher estimated coefficients for the standardized variables. Among the other top nine features, the number of rooms, the number of persons, household income, building structure types, family life cycle, heating fuel type, and tenure types are available in the ACS data. Therefore, these household-level variables are all controlled in the synthesizing process. The TAZ level residential operation energy consumption is then calculated by aggregating the consumptions of each synthesized household in by TAZ.

5. Results

The results for electricity energy consumption suggests each resident consumes approximately 19.43 million BTU (or 5694 KWh) per year in the 10-county Atlanta metropolitan area. The electricity consumption per capita for the State of Georgia is estimated as 19.56 million BTU per year (EIA, 2009). The results show that residents located in the central metro and peripheral areas tend to consume more electricity. The results for natural gas energy indicates the average consumption per person is approximately 11.25 million BTU per year, which is close to Georgia's consumption per capita (12.6 million BTU per year). The natural gas consumptions decrease in peripheral areas, where the heating fuel is primarily electricity or other fuel sources. The results for other energy consumptions show the peripheral residents tend to consume more energy generated by other fuel types.

6. Conclusions

In this study, we developed an energy synthesizing model, which can be potentially applied to any major U.S. cities. Most bottom-up energy estimation models are data intensive and, therefore, constrained by the availability of micro-sampled data and regional housing inventories. To address these model limitations, we

developed a novel estimation technique comprising of statistical matching, machine learning and population synthesizing. The developed model requires data such as RECS, PUMS, and ACS, which are available across the nation. We used data from 10-county Atlanta metropolitan area with 1593 zones to test our estimation process. The results show that the electricity, natural gas, and other energy uses per capita are 19.17, 14.05, and 0.59 million BTU per year, respectively. These results are consistent with statistics for the State of Georgia, which are 19.56 and 12.6 million BTU per year for electricity and natural gas consumptions for the same year. Additionally, the synthesized results are also cross-compared with the electricity and natural gas utility bills available for 30 zip codes. The correlations between our results and observed consumption are 0.908 and 0.927 for electricity and natural gas uses. Given the validation results we can conclude that this approach offers an exciting new method to estimate residential energy consumption in U.S. metros.

References

Aydinalp, M., Ugursal, V. I., & Fung, A. S. (2002). Modeling of the appliance, lighting, and space-cooling energy consumptions in the residential sector using neural networks. Applied Energy, 71(2), 87–110.

Aydinalp, M., Ugursal, V. I., & Fung, A. S. (2004). Modeling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks. Applied Energy, 79(2), 159–178.

Chen, J., Wang, X., & Steemers, K. (2013). A statistical analysis of a residential energy consumption survey study in Hangzhou, China. Energy and Buildings, 66, 193–202.

Crawley, D. B., Hand, J. W., Kummert, M., & Griffith, B. T. (2008). Contrasting the capabilities of building energy performance simulation programs. Building and Environment, 43(4), 661–673.

de Normalización, C. E. (2008). EN ISO 13790: Energy Performance of Buildings: Calculation of Energy Use for Space Heating and Cooling (ISO 13790: 2008). CEN.

Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical region. Energy and Buildings, 37(5), 545–553.

Douthitt, R. A. (1989). An economic analysis of the demand for residential space heating fuel in Canada. Energy, 14(4), 187–197.

Fung, A. S.-L., Aydinalp, M., & Ugursal, V. I. (1999). Econometric models for major residential energy end-uses. CREEDAC, Dalhousie University.

Haas, R., & Schipper, L. (1998). Residential energy demand in OECD-countries and the role of irreversible efficiency improvements. Energy Economics, 20(4), 421–442.

Hirst, E., Lin, W., & Cope, J. (1977). Residential energy use model sensitive to demographic, economic, and technological factors. Q. Rev. Econ. Bus.;(United States), 17(2).

Hirst, Eric. (1978). A model of residential energy use. Simulation, 30(3), 69–74.

Hirst, Eric, Goeltz, R., & White, D. (1986). Determination of household energy using “fingerprints” from energy billing data. International Journal of Energy Research, 10(4), 393–405.

Jain, R. K., Smith, K. M., Culligan, P. J., & Taylor, J. E. (2014). Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Applied Energy, 123, 168–178.

Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, A., Stevanovic, Z., & Djurovic-Petrovic, M. (2010). A review of bottom-up building stock models for energy consumption in the residential sector. Building and Environment, 45(7), 1683–1697.

Nesbakken, R. (1999). Price sensitivity of residential energy consumption in Norway. Energy Economics, 21(6), 493–515.

Raffio, G., Isambert, O., Mertz, G., Schreier, C., & Kissock, K. (2007). Targeting residential energy assistance. In ASME 2007 Energy Sustainability Conference (pp. 489–495). American Society of Mechanical Engineers.

Saha, G. P., & Stephenson, J. (1980). A model of residential energy use in New Zealand. Energy, 5(2), 167–175.

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. Renewable and Sustainable Energy Reviews, 13(8), 1819–1835.

Tonn, B. E., & White, D. L. (1988). Residential electricity use, wood use, and indoor temperature; An econometric model. Energy Systems and Policy;(USA), 12(3).

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., & Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In 88th Annual Meeting of the Transportation Research Board, Washington, DC.

Zhang, Q. (2004). Residential energy consumption in China and its comparison with Japan, Canada, and USA. Energy and Buildings, 36(12), 1217–1225.

Zhao, H., & Magoulès, F. (2012). A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews, 16(6), 3586–3592.