Estimating Residential Energy Consumption in Metropolitan Areas: A Microsimulation Approach

Wenwen Zhang[a*], Caleb Robinson[b], Subhrajit Guhathakurta[c], Venu M. Garikapati[d], Bistra Dilkina[e], Marilyn A. Brown[f], and Ram M. Pendyala[g]

[a] Virginia Polytechnic Institute and State University, Department of Urban Affairs and Planning, Blacksburg, VA
[b] Georgia Institute of Technology, School of Computational Science & Engineering, Atlanta, GA
[c] Georgia Institute of Technology, School of City & Regional Planning, Atlanta, GA
[d] National Renewable Energy Laboratory, Golden, CO
[e] University of Southern California, Department of Computer Science, Los Angeles, CA
[f] Georgia Institute of Technology, School of Public Policy, Atlanta, GA
[g] Arizona State University, School of Sustainable Engineering and the Built Environment, Tempe, AZ

Highlights

Machine learning algorithms can address energy consumption data gaps
Residential Energy Consumption Survey matched with synthesized households to generate neighborhood energy footprints
Validation uses zip code power consumption data provided by the energy provider

Abstract

Prior research has shown that land use patterns and the spatial configurations of cities have a significant impact on residential energy demand. Given the pressing issues surrounding energy security and climate change, there is renewed interest in developing and retrofitting cities to make them more energy efficient. Yet deriving micro-scale residential energy footprints of metropolitan areas is challenging because high resolution data from energy providers is generally unavailable. In this study, a bottom-up model is proposed to estimate residential energy demand using datasets that are commonly available in the United States. The model applies novel machine learning methods to match records in the Residential Energy Consumption Survey with Public Use Microdata samples. This matching and machine learning produce a synthetic household energy distribution at a neighborhood scale. The model was tested and validated with data from the Atlanta metropolitan region to demonstrate its application and promise.

Keywords: Residential Energy Consumption, Data Synthesis, Statistical Matching, Machine Learning

[*] Corresponding author at: School of Urban Affairs and Planning, Virginia Polytechnic Institute and State University, Blacksburg, VA
E-mail address: wenwenz3@vt.edu

1. Introduction

The residential sector consumes a little more than one-fifth of total energy demand in the U.S., and this proportion has not changed significantly for about 30 years. However, during this period the sector's consumption grew by around 28%, which was slightly slower than the growth in population (30%) and significantly slower than the growth in the number of homes (40%) (U.S. Energy Information Administration, 2016). The slower pace of growth in energy consumption within homes has been attributed to better building design, construction, and retrofitting with improved products such as low-e windows and foam insulation. The efficiency of lighting, home appliances and heating, ventilation, and air conditioning (HVAC) systems has also advanced (Brown & Wang, 2015).

While the slower growth in demand for household energy has been a welcome trend for energy and climate security, there is concern that further progress will be increasingly challenging without new initiatives. The average U.S. household now uses many more consumer electronics – in particular, computers, cell phones, tablets, and related devices. In addition, the average house has added 1000 square feet since 1973 and this trend is continuing. Therefore, new strategies are being debated to further shrink U.S. reliance on fossil fuels. One such strategy involves designing and retrofitting urban areas to reduce their energy requirements and to offer opportunities for shorter travel distances. For instance, one study compared the impact of residential development density on residential energy use in Toronto and suggested that compact developed pattern should be encouraged to reduce residential energy consumption (Norman, MacLean, & Kennedy, 2006). An Austin, TX-based empirical study also found that life-cycle residential and transportation energy demands for households living in densely developed areas are significantly lower compared with their peers in suburban areas (Nichols & Kockelman, 2014). A more recent study also arrived at a similar conclusion using data from Phoenix, AZ (Guhathakurta & Williams, 2015). The link between density and energy footprints has also corroborated by statistical analysis of the 100 largest metropolitan areas in the U.S. (Brown, Southworth, & Sarzynski, 2009).

Although prior research has indicated that urban form is significantly related to household operation and transportation energy use, most studies have focused on analyzing and modeling energy savings in the transportation sector. W. Zhang, Guhathakurta, & Ross (2016)'s study in Phoenix, AZ, suggested that more compact development in suburban areas can reduce travel energy footprint. Garikapati et al. (2017) developed a generalizable method to estimate travel energy consumption based on built environment features and outputs from regional transportation demand model. The studies on the impact of urban form/neighborhood design on residential operational energy use, is constrained by a lack of residential energy consumption data at fine spatial granularity (Yekang Ko, 2013). Determining the energy footprint of the residential sector in a metropolitan region at the neighborhood scale is challenging, largely because information on household energy use is difficult to obtain from utility companies due to legal and privacy concerns. In the absence of such objective data, household surveys are utilized to collect information about energy use together with the factors that drive household energy consumption, such as the type of housing, occupant behaviors, type and number of appliances and devices, and built environment characteristics. Such surveys are time-consuming and expensive especially if representative samples are to be drawn at the neighborhood scale. The most authoritative and detailed survey of residential energy use, which is known as the

Residential Energy Consumption Survey (RECS), is conducted by the U.S. Energy Information Administration (EIA). The sample of households in RECS is nationally representative; however, subsamples compiled using the geographic identifiers available in RECS may not be representative of more disaggregate residential neighborhoods under the geographic identifiers. Moreover, household records in RECS are not available for small geographic units, such as neighborhoods. Therefore, understanding how the design of neighborhoods will influence residential energy consumption is difficult from exclusively RECS data.

Developing a robust disaggregated residential consumption model is useful for multiple planning purposes. Model predictions can be used to determine the variation in city or regional energy demand, especially for comparing different newly developed areas or neighborhoods that experience significant demographic changes. Model coefficients can be interpreted to identify the sensitivity of residential energy consumption to property characteristics, user attributes, and urban form parameters such as density and land use diversity. Model results can be used to devise policies and provide guidance for designing energy efficient neighborhoods. The model may also be used to assess residential energy consumption under various scenarios of market penetration of new technology, energy prices, and urban form changes. This study addresses the objectives listed above by developing a novel bottom-up technique for generating estimates of neighborhood level residential energy consumption by: 1) matching RECS data with other publicly available datasets using machine-learning algorithms; and 2) synthesizing residential energy consumption at the level of traffic analysis zones for the study region. The methodology not only overcomes many of the challenges of existing modeling approaches but is also easily replicable for all U.S. metropolitan areas.

The residential energy estimation technique described in this paper is tested and validated using data from the Atlanta metropolitan area. Given that the input data used for the Atlanta metro case study are all widely available for most other U.S. metropolitan regions, the proposed methodology can be widely applied to other large cities in the U.S. as well. To enable wide application of the developed tool, the code is open sourced and available through a public GitHub repository (available at https://github.com/SEI-ENERGY). Additionally, other researchers can further improve these tools and advance this line of research.

The rest of the paper is divided into five sections. Section two outlines the prior foundational research that our current work is now advancing. The third section describes the various methodological approaches that enabled our work. Section four demonstrates the application of the methods using data from the Atlanta metropolitan region. Section five offers a discussion of the results from the neighborhood level residential energy estimates derived in section four. Finally, section six presents summary remarks and limitations of the approach.

2. Prior Studies

The literature on modeling and estimating residential energy consumption can be generally classified into two types: 1) top-down models and 2) bottom-up models. Top-down models typically examine regional estimates of residential energy consumption using longitudinal aggregated data about fairly macro-scale factors, such as gross domestic product (GDP), economic activities, and population. Bianco, Scarpa, & Tagliafico (2014) found that

residential natural gas consumption in Italy is linearly correlated with GDP, price of the fuel and population size. The price of energy and the adoption of energy equipment have also been positively correlated with residential energy consumption using data from Italy (Ghosh & Kanjilal, 2014), along with the aggregate age, evolution, and density of housing. Glaeser & Kahn (2010) showed that aggregated residential energy consumption/emission is higher in U.S. regions where overall housing density is comparatively lower and such phenomenon is typically significant in older regions older regions, such as New York . A longitudinal study in China also suggested that residential energy consumption at the province level is negatively correlated with the urbanization process (Wang, Fang, Guan, Pang, & Ma, 2014). County-level data are often used to characterize the impact of urban form on residential energy consumption at the metro scale. For instance, Brown & Cox (2015) estimated energy consumption and carbon emissions for counties in the 100 largest U.S. metropolitan areas. These models are often used to assess total the total energy demand of cities and metropolitan areas, but they are not suitable for evaluating variations in energy consumption across neighborhoods. Top-down models are therefore ineffective in determining the technological upgrades or urban built environment reconfigurations that would help reduce residential energy footprint at the neighborhood scale.

The bottom-up models, in contrast, estimate residential energy consumption using micro-samples collected from the region. The bottom-up approach can be implemented through either statistical models or engineering models. The statistical models relate unique contributions of various factors to residential energy consumption. Raffio, Isambert, Mertz, Schreier, & Kissock (2007) linearly associated housing units, socio-economic, and demographic features of households and weather data with residential energy profile. Fung, Aydinalp, & Ugursal (1999) showed that residential energy consumption is linearly correlated with energy price, household demographics, appliance features, and weather, across different end-uses. Lopes, Antunes, & Martins (2015) associated household electricity consumption with household behaviors, such as weekly washes, appliance use knowledge, and other control variables, such as income, education, and demographics. Their results showed that electricity consumption is positively related with intensity of use and individual's knowledge regarding how to adjust appliances for more efficient energy use. As with county-level heating and cooling degree days in top-down models, urban variations in micro-climates including heating and cooling degree days are often included in bottom-up models (Kavousian, Rajagopal, & Fischer, 2013). The principal aim of statistical models is to help inform the design of policies that influence energy prices (Chen, Wang, & Steemers, 2013) and the formulation of incentives to promote energy conservation by encouraging changes in user behaviors (Tso & Guan, 2014). Historically, statistical models have seldom been used to assess total energy demand in a region. Some recent studies, however, have developed advanced machine learning-based statistical models to predict residential energy consumption. Aydinalp, Ugursal, & Fung used neural networks to predict appliance, lighting, and space-cooling energy consumptions (2002) and domestic hot-water heating energy consumption (2004) in residential buildings and found that neural networks can significantly improve the predictability. Dong, Cao, & Lee (2005) applied support vector machine models to predict building energy consumption in tropical regions. Jain, Smith, Culligan, & Taylor (2014) also successfully trained support vector machine models to predict energy consumption in multifamily housing units. However, these models are rather data intensive and difficult to apply for large study areas.

Engineering models compute energy consumption based on energy ratings of various appliances, building materials, and energy saving technologies on-site using thermodynamic theorems (Zhao & Magoulès, 2012). Specifically, engineering models first estimate energy consumption for a series of typical prototypes or archetypes of housing stocks in the region, using a small sample of buildings (Hargreaves, Cheng, Deshmukh, Leach, & Steemers, 2017). The prototypical units for which energy consumption is derived are then compared to their prevalence in the region and aggregated to derive the total residential energy footprint. The objective here 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 from the prototypical housing stocks to the region is usually accomplished by assigning weights to the sampled buildings based on the regional housing inventory. A critical drawback of these engineering models is that socio-economic characteristics and occupant behaviors are not captured but instead are simplified using various assumptions. Many calculations, codified in software systems, have been developed to estimate building energy consumption using this modeling approach. Crawley, Hand, Kummert, & Griffith (2008) conducted a comprehensive comparison of twenty programs and the results suggest that the required model inputs are quite large and available only at regional scales. The International Organization for Standardization (ISO) guidebook for building energy consumption estimation also indicate that engineering models require extensive microscopic data inputs (de Normalización, 2008).

Despite different model objectives, the statistical and engineering models share some limitations. First, both model approaches rely heavily on the availability of historical micro-level sample data, which are usually time-consuming and expensive to collect. Swan & Ugursal (2009) reviewed many residential energy prediction studies and pointed out the lack of micro-level data as a bottleneck for many modeling efforts. Similar issues are also identified in a more recent building energy modeling review effort by Zhao & Magoulès (2012). For instance, as noted earlier, the U.S. EIA publishes RECS every three to five years but the sample size in each metropolitan area is quite small, rendering it insufficient to assess local energy consumption. Additionally, it is difficult, if not impossible, to extrapolate the building-level model results to the region, as the features considered in the models are usually not available across the region (Kavgic et al., 2010).

To address the limitations of previous approaches for residential energy estimation, this study proposes a new strategy for bottom-up estimation of residential energy consumption. The proposed approach uses Residential Energy Consumption Survey (RECS), Public Use Microdata Sample (PUMS), and American Community Survey (ACS) data as inputs and provides synthesized households with appended energy consumption for the study region as outputs. The input data are available for all major metropolitan areas in the U.S. Therefore, the methodology presented in this paper can potentially be applied to estimate residential energy consumption in any U.S. metro region.

3. Methods

The neighborhood-level residential energy consumption estimates are derived in four sequential steps: 1) statistically match household records between the RECS and PUMS data, 2) estimate energy consumption models using the matched records and impute energy consumption for the unmatched records in PUMS data, 3) synthesize households using enhanced PUMS and ACS data, and 4) aggregate the energy consumption estimates by households to the level of defined subareas, such as traffic analysis zones (TAZs).

The objective of the first two components of the modeling framework is to develop energy consumption models using explanatory variables available in the PUMS data so that residential energy consumption can be estimated for households in the regional synthetic population. The PUMS data, however, do not contain energy consumption information which is critical for developing energy source specific models. The RECS data, on the other hand, do have comprehensive energy use variables by energy source for each sampled household. Yet, RECS cannot be used to develop metropolitan-scale models since its sample sizes are too small and not representative. More importantly, the sample households in the RECS data are not geo-located. Therefore, our proposed model first statistically matches households in RECS with PUMS data based on variables that are common to both data sets. In other words, a statistical matching process is employed to "join" the energy consumption variables from the RECS data to a portion of the PUMS data based on the similarity of variables present in both data sets. In the second step, a set of models is estimated using energy consumption variables (merged 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 models are then applied to the unmatched PUMS data to impute energy consumption for these records. At the end of the first two steps described above, all the PUMS data records are appended with energy consumption information by source for each sample household.

The third step takes the PUMS data with added energy consumption information as the seed matrix and relevant ACS data as the marginal controls to synthesize a complete population of households in the study region. Both household-level and population-level variables, which are highly correlated with energy consumption in the estimated models, are controlled to obtain a more representative profile. Finally, the synthesized energy consumption at the household level is aggregated up to selected geographic units to estimate the spatial distribution of energy consumption in the region. Because of the novelty of this machine learning approach, the following sections elaborate on our modeling steps.

3.1 Statistical Matching

Statistical matching methods are widely used when no single dataset has the full set of variables needed for further analysis. In this study, both residential energy consumption estimates and a wide range of household and population level socio-economic variables are needed to be present in the same dataset to estimate residential energy consumption as a function of household, person, and housing unit attributes. Since the RECS data comprises small sample sizes at metropolitan area scale, it is first statistically matched with PUMS data to append residential energy consumption variables to a set of statistically similar households in the PUMS data.

This statistical matching framework begins with RECS data and PUMS data sharing a limited set of variables, X, while residential electricity, natural gas, and other energy consumption variables, Z, are only available in the RECS data and many other socio-demographic and economic variables, Y, are only available in the PUMS data. The objective of this data matching step is to join two datasets using X so that Y and Z can be present in the same table (i.e., $RECS \cup PUMS$). After identifying the common variables X, it is necessary to perform some "harmonizing" procedures to ensure that the common variables share the same measurement criteria. For instance, ordinal income variables may have different income ranges for each bin, necessitating some adjustment to facilitate the matching process. Additionally, some categorical variables, such as heating fuel types also need to be cross-compared to confirm the match in the definitions of classification methods. Finally, the common numerical variables are standardized to z-scores using mean, μ_{pooled}, and standard deviation, σ_{pooled}, calculated with X from both RECS and PUMS to eliminate the impacts of alternative measurement units in the matching process. The formula is shown below:

$$z_i = \frac{\left(x_i - \mu_{pooled}\right)}{\sigma_{pooled}} \tag{1}$$

Furthermore, not all X can be used for matching purposes; only the X_M ($X_M \subseteq X$) that is most relevant to Z should be used as the matching variables (D'Orazio, 2016). The correlations between X and Z variables are estimated using Spearman's rank correlation coefficients. After determining the matching variables that are strongly correlated with energy consumption, X_M, the two datasets are joined using the nearest neighbor distance hot deck micro-matching method provided in the R package StatMatch. This a nonparametric approach, which is frequently used in statistical matching when the objective is to generate a synthetic dataset with X, Y, and Z variables (D'Orazio, 2016). This method searches in the PUMS data for the nearest neighbor of each record in the RECS data according to the distance calculated on the continuous matching variables. The distance, d, is estimated as the Manhattan Distance (i.e., the default measurement in StatMatch Package), as shown in the equation below. Manhattan Distance is sufficient to perform one-to-one matching between RECS and PUMS data, as there is a small number of continuous matching variables and the variables are standardized before matching. The Manhattan Distance is also selected because it can reduce computation costs of the matching process

$$d = \sum_{i=1}^{n} |X_{M_i}^{RECS} - X_{M_i}^{PUMS}| \tag{2}$$

Where,
n, is the total number of matching variables;
$X_{M_i}^{RECS}$, refers to the i[th] matching variable from the RECS dataset;
$X_{M_i}^{PUMS}$, refers to the i[th] matching variable from the PUMS dataset.

The ordinal and nominal matching variables are used to define matching classes. Only records in the same matching classes can be matched together, i.e., only records that fall into the same income category, heating fuel type, etc. can be matched with each other. The matching quality is then evaluated by 1) examining the distribution of matching distances between records and 2) comparing the correlation structure of Z and X_M variables in the RECS data and the matched data (Rässler, 2012).

3.2 Residential Energy Consumption Imputation

Various machine learning models are used to impute the residential energy consumption for the unmatched PUMS records across different energy sources. The models are trained using the set of matched PUMS records. The target features are residential energy consumption by BTU (British Thermal Unit) type. All other socio-demographic, economic and housing unit variables in the PUMS data are used as explanatory features. The categorical variables are transformed into sequences of binary variables before including them in the models. Variables that are considered irrelevant to energy consumption estimation are excluded manually. To prevent over-fitting and improve the generalizability of the model to other datasets, the 10-fold cross validation method is used to select the best models and parameters. The examined models include Linear Regression, Ridge Regression (Hoerl & Kennard, 1970), Lasso Regression (Tibshirani, 1996), Elastic Net Regression, Bagging (Breiman, 1996), Random Forest (Liaw & Wiener, 2002), Support Vector Machine (SVM) (Smola & Schölkopf, 2004), AdaBoost (Collins, Schapire, & Singer, 2002), Gradient Boosting (Friedman, 2002), and Extra Trees. The Mean Absolute Percentage Errors (MAPE), Mean Absolute Errors (MAE), and Average R^2 score on the testing datasets are calculated to evaluate and compare the overall performance of the models. For some households, the observed consumption for natural gas and other source of energy is zero, rendering it impossible to some models with MAPE scores. Therefore, the models with the lowest MAE values are used to impute source specific energy consumption for unmatched PUMS records. The model development, parameter estimation, cross-validation, and algorithm application procedures are implemented using 2.7 Python Scikit-learn library (Pedregosa et al., 2011).

3.3 Household Synthesis

The final PUMS data (with imputed energy consumption) output from the previous step, together with ACS data on the marginal distributions of various socio-economic attributes of the population, are then used to generate a synthetic population of households for the entire metropolitan region. The PUMS records, with estimated energy consumption variables, are essentially replicated (weighted and expanded) to produce a synthetic population of household agents in the region that mimics the distributions of socio-economic attributes as depicted in the ACS data. The PUMS data provide joint distributions for various socio-economic and demographic variables. The ACS data provide marginal distributions for various socio-economic and demographic variables at census tract, block group and block levels. In this study, the synthetic population is generated such that population distributions are matched at the block group level, thus ensuring that the synthetic population is highly representative of the true population in the region. The block group level synthetic population records can be aggregated to any desired level of geographic resolution (say, traffic analysis zone, regional planning district, or zip code) to perform neighborhood level analysis of consumption patterns. The Iterative Proportional Updating (IPU) algorithm embedded within the PopGen software package is used to

ensure that both household and person characteristics are controlled and matched in the population synthesis process (Konduri, You, Garikapati, & Pendyala, 2016).

4. Data and Model Implementation

The modeling approach developed in the previous section is tested by applying it to the 10-county Atlanta metropolitan area for generating TAZ-level residential energy consumption estimates. The results are then compared with electricity and natural gas consumption data provided by Georgia Power and Atlanta Gas Light to validate the model outputs.

4.1. Data

Data used to implement our proposed model include the 2009 RECS public use micro-dataset, the 2009 PUMS data, and the 2009 ACS data. The RECS data contain information about various attributes of housing units and some socio-economic and demographic characteristics of the occupants. The latest RECS data with household-level energy consumption and energy expenditures are available for the year 2009; the data contain 2,246 sampled households in the southeast region (i.e., Subdivision 5). Because the analysis is being done for the Atlanta metropolitan region, only the RECS records from the southeast region were used for energy consumption model estimation with a view to control for geographic disparities in energy consumption patterns across the country.

The PUMS data include housing unit and population records with individual response information collected from the American Community Survey (ACS). In contrast to RECS data, the PUMS data contain significantly more information on demographic and socio-economic attributes and contain a limited set of housing unit level features. While energy consumption data are not included, the PUMS data do contain self-reported energy bills (expenses) as a proxy measure of residential energy consumption. Compared with RECS data, PUMS data provide a larger sample size for the State of Georgia. There are 37,009 household records in the 2009 Georgia PUMS data.

The ACS data provide summary statistical information about housing units, households, and persons on a yearly basis by the designated census geographic units, such as census tracts, block groups or blocks. The data serve as an authoritative source of information on the marginal distributions of specific characteristics of the population residing in the defined geographic units. Most attributes present in the PUMS data have corresponding marginal distributions available in the ACS data. Therefore, this dataset is used to derive marginal controls necessary to generate representative synthetic households.

4.2. Statistical Matching Implementation

RECS data in Region 5 (i.e., the Southeast region) are statistically matched with PUMS data for 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, the identify shared variables are tabulated in Appendix A. There are 12 common variables, including the type of housing unit, property ownership, year built, energy bills in the two datasets. After close examination, it is noticed that although these variables measure the same aspect of households or housing units, the measurement units or categories used can vary significantly. To harmonize the measurements,

the common variables are reclassified or reorganized through appropriate transformations or aggregation and reconciliation of categories.

The Spearman rho rank correlation analysis is conducted 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 are therefore used as group control variables, indicating that only households with the same household structure and 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 some housing categories with zero observations if all heating fuel types are retained 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 in the PUMS will not be matched and appended with energy consumption. Therefore, to simplify the matching process and ensure a high level of statistical data fusion, the heating fuel types are reclassified into three categories: 1) electricity, 2) natural gas, and 3) other fuels.

The continuous variables, such as the total number of rooms, and annual electricity, natural gas, and other fuel expenses are used as match variables. To eliminate the influence of differing measurement units on the Manhattan Distance calculation outcome, all of 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 for joining ELBTU, given two reasons: 1) it is highly correlated with the total number of rooms in both datasets and 2) there are many missing bedroom values in the PUMS dataset. There are no missing values for the selected matching variables.

Table 1: Spearman rho rank correlation analysis results

Variable names		$\boldsymbol{Adjusted\ \rho^2\ [p\ value]}$			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 within the same category as defined by the group control variables. For each RECS record, the closest PUMS record is found and matched. For ELBTU, the maximum estimated distance between the matched records is 14.11. This is because there is an observation with 23 rooms and $19,040 in annual electricity expenses in the RECS data; the closest household that the algorithm found in the PUMS data is a household with 17 rooms and $6,480 in annual electricity expenditures. Including the above outlier, there are only three matched pairs with a distance larger than 1. To ensure that only households that closely resemble one another are matched together, these three outlier households are removed from the final matched outputs. The final median distance is just 0.038, and the 75th percentile distance value is 0.057, indicating that the data are well matched together (Rässler, 2012). The maximum distance for matching results of OBTU is 33.45. This is because one household in the RECS data has an annual bill of $12,972 for other energy sources, while the household that consumed the most in other fuels in the PUMS data had an expenditure of $4,300 annually. Therefore, these two households are not compatible from a statistical matching perspective and are removed from the output. The maximum distance for the rest of the matched records is 0.98. For the NGBTU matching results, the maximum distance is 0.06, suggesting all of the households are matched successfully.

To validate the matching results, ordinary least squares (OLS) regressions are estimated, with ELBTU, NGBTU, and OBTU as the dependent variables and the corresponding categorical and continuous matching variables from both the RECS and PUMS data as the independent variables. The detailed results are summarized in Table 3. The adjusted R-square of the models are reasonably high, and the significance, signs, and magnitude of estimated coefficients are consistent across models using the RECS and PUMS data. In sum, all of these outputs suggest the matching results are robust.

Table 3: OLS Models using variables from RECS and PUMS by BTU Types

Variables	**ELBTU**		**NGBTU**		**OBTU**	
	RECS	**PUMS**	**RECS**	**PUMS**	**RECS**	**PUMS**
(Intercept)	-1418.7*	-1498.2*	960.9*	1050.8*	-6.7	-5.6
	(-2.007)	(-2.104)	(2.315)	(2.492)	(-0.044)	(-0.037)

Total Rooms	1063.5*** (10.024)	1047.9* (9.779)	-	-	-	-
Electricity Bill	25.3 *** (102.337)	25.4 *** (101.829)	-	-	-	-
Natural Gas Bill	-	-	73.6*** (170.681)	73.5*** (167.753)	-	-
Other Bill	-	-	-	-	42.6*** (166.426)	42.4*** (166.142)
Heat Fuel - Electricity	1978.3*** (4.285)	1973.8*** (4.257)	-1434.7*** (-3.340)	-1496.0*** (-3.428)	9.2 (0.051)	18.2 (0.100)
Heat Fuel - Other	3386.9 (3.418)	3495.1 (3.511)	-1121.4 (-1.445)	-1162.5 (-1.475)	-4180.7*** (-9.253)	-4085.8*** (-9.033)
Adjusted R^2	0.88	0.88	0.96	0.96	0.94	0.94
N	2243	2243	2246	2246	2245	2245

4.3. Energy Consumption (BTU) Imputation Model Results

The features in the matched PUMS dataset are first processed. Variables with more than 10% missing values are not considered in the models. This excludes 28 variables in the PUMS dataset. Among the remaining variables, nine continuous variables are standardized. The remaining 36 categorical variables are converted into 134 binary variables. The total number of variables included in the model is 143. The averaged results of the 10-fold cross-validation experiments are shown in Table 4 through 6. All of the training models use the default parameter settings in the Scikit-learn package. The results are sorted by Mean Absolute Errors.

The outputs for electricity consumption suggest that the Elastic Net regression performs the best among all tested models, as shown in Table 4. The Elastic Net model presents the smallest mean and median absolute errors and the smallest average percent difference. The results suggest, on average, that predicted consumption is approximately 13.4% different from the statistically matched electricity BTU consumption values. Additionally, the model has the largest average $\mathbf{R^2}$ among all of the examined models. The results also suggest that other linear models, such as Lasso, Ridge, and Ordinary Least Squares also show similar prediction power (in terms of the magnitude of errors and $\mathbf{R^2}$), indicating that the relationship between explanatory variables and the target variable (ELBTU) may be considered linear. In contrast, ensemble learning models, such as Bagging, Random Forest, Gradient Boosting, AdaBoost, and Extra Trees, perform comparatively poorly, with higher absolute errors and lower $\mathbf{R^2}$ values.

Table 4: Cross-Validation Results for Electricity Energy Consumption Models

Models	**Mean Absolute Error**	**Median Absolute Error**	**R^2**	**Mean Absolute Percentage Error**
Elastic Net	**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

Table 5 shows experimental results for prediction of natural gas energy consumption. The mean average percent difference is not reported, given that many households do not use natural gas, rendering many cases with zero denominator. The results show that there are tradeoffs between models. For instance, Elastic Net models have the smallest Mean Absolute Errors and highest $\mathbf{R^2}$. The Bagging models, on the other hand, have the lowest Median Absolute Errors, with more than half of the predictions with zero errors. The performance of Random Forest models is also impressive, with the second-best results across all metrics. Because it has the smallest Mean Absolute Error, the Elastic Net model is used to impute missing NGBTU values for the unmatched PUMS records.

Table 5: Cross-Validation Results for Natural Gas Energy Consumption Models

Models	Mean Absolute Error	Median Absolute Error	R^2
Elastic Net	**2.88e+03 +/- 318.263**	314.501 +/- 77.475	**0.959 +/- 0.006**
Random Forest	3.07e+03 +/- 338.450	79.895 +/- 11.292	0.949 +/- 0.008
Lasso	3.13e+03 +/- 295.669	627.051 +/- 115.948	0.958 +/- 0.006
Ridge	3.13e+03 +/- 295.737	637.697 +/- 113.990	0.958 +/- 0.006
Linear	3.15e+03 +/- 297.046	657.148 +/- 111.088	0.957 +/- 0.006
Bagging	3.18e+03 +/- 294.971	**0.000 +/- 0.000**	0.943 +/- 0.007
Gradient Boosting	4.71e+03 +/- 317.574	2.20e+03 +/- 51.461	0.939 +/- 0.007
Extra Trees	5.18e+03 +/- 667.890	756.787 +/- 184.247	0.902 +/- 0.021
AdaBoost	5.69e+03 +/- 1.78e+03	3.80e+03 +/- 2.40e+03	0.938 +/- 0.017

The model outputs for other BTU consumption, as shown in Table 6, suggest that Random Forest models perform the best with the lowest Mean Absolute Errors of approximately 834.6. The Bagging models again have the lowest median absolute errors at 0. Elastic Net, on the other hand, provides the highest R^2 at 0.93. Similar to natural gas models, the percent differences are not reported as some observations have zero consumption.

Table 6: Cross-Validation Results for Other Energy Consumption Models

Models	Mean Absolute Error	Median Absolute Error	R^2
Random Forest	**834.556 +/- 138.063**	13.111 +/- 1.995	0.918 +/- 0.022
Bagging	882.092 +/- 173.000	**0.000 +/- 0.000**	0.900 +/- 0.033
Elastic Net	981.025 +/- 121.736	32.673 +/- 37.458	**0.930 +/- 0.017**
Lasso	1.19e+03 +/- 98.960	197.003 +/- 60.443	0.927 +/- 0.021
Ridge	1.20e+03 +/- 101.890	196.339 +/- 61.706	0.927 +/- 0.021
Linear	1.22e+03 +/- 103.395	206.223 +/- 66.615	0.926 +/- 0.022
Gradient Boosting	1.30e+03 +/- 124.141	476.773 +/- 17.068	0.905 +/- 0.031
AdaBoost	1.64e+03 +/- 542.502	857.917 +/- 652.291	0.910 +/- 0.023
Extra Trees	1.67e+03 +/- 241.021	740.746 +/- 261.793	0.894 +/- 0.026

In summary, the cross-validation results show that models using features available in the PUMS data perform quite well in capturing the variations in energy consumption. The R^2 values of the best models for each fuel source are all close to or above 0.90. The trained Elastic Net regressors are used to impute ELBTU and NGBTU, and the Random Forest regressor is used to impute OBTU. The above trained models are then applied to unmatched Georgia PUMS records. The final output from this model component is an enhanced complete PUMS dataset that has three appended columns, namely, matched or imputed ELBTU, NGBTU, and OBTU values.

4.4. Synthetic Population Generation

Synthetic population generation is accomplished through the use of the PopGen 1.1 software package. This software uses PUMS data as the seed matrix comprising joint distributions among various features of households, and ACS data as the source of marginal controls, to estimate a weight for each household in the PUMS data. These weights provide a basis for expanding the PUMS data into a full synthetic population for the region. Both household (or housing unit) level and person level variables can be controlled and matched in the synthesis process through the application of the iterative proportional updating (IPU) algorithm embedded within PopGen. The control variables are selected based on correlations with the target variables of interest (i.e., ELBTU, NGBTU, and OBTU) and the availability of information in the ACS data.

The correlation of various explanatory variables with target variables of interest is determined using estimated coefficients from the Elastic Net models and the feature importance scores from the Random Forest model. The estimated top 10 coefficients for electricity and natural gas consumption Elastic Net models are shown in Table 7. The results show that annual electricity and natural gas expenditures are highly correlated, as expected, with the electricity and natural gas BTU consumption, with significantly higher estimated coefficients for the standardized variables. However, these variables cannot be used as marginal controls in the synthesis process due to their absence in the 2009 ACS data. 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 (ownership/rent) types are available in the ACS data. Therefore, these household-level variables are all controlled in the synthesis process.

Table 7: Top 10 Most Important Features based on Elastic Net Model Coefficients

Electricity BTU Model			Natural Gas BTU Model		
Features	**Feature Descriptions**	**Coef.**	**Feature**	**Feature Descriptions**	**Coef.**
ELEP	Electric (Yearly Cost)	14275.7	GASP	Gas (Yearly Cost)	19781.3
RMSP*	Number of Rooms	2288.1	HFL_1*	Heating Fuel Type (Gas)	4550.2
BDSP*	Number of Bedrooms	1539.8	WATP	Water (Yearly cost)	1386.9
NP*	Number of Person	1288.8	HINCP*	Household Income	851.5
BLD_2*	Single Family Detached	989.3	RMSP*	Number of Rooms	661.5
HHT_1*	Married Couple Family	790.9	BDSP*	Number of Bedrooms	566.2
HFL_3*	Heating Fuel Type (Elec.)	618.4	SVAL_1	Specified value owner unit	560.9
VEH_3*	Three Vehicles Available	565.7	YBL_6*	Structure built in 1980s	394.5
WATP	Water (Yearly cost)	562.4	NP*	Number of Person	331.6
SRNT_0	Not Specified Rent Unit	486.1	TEN_1*	Property Owned with Mortgage	276.6

* features available in the 2009 ACS data

The importance score for "other fuel expenditure", among features used in the Random Forest regressor, is significantly higher than for all other features. The importance score for other fuel expenses is 0.95. The second most important feature is the household water utility expenses. However, the number of rooms, the number of persons, household income, building structure types, heating fuel type, and tenure type are also in the top 10 important features list. Although various utility expenses are not controlled in the synthesis process, other critical features for which information is available in the ACS data are all controlled. Therefore, the synthesis process provides a representative population with a distribution of energy consumption by source that reflects energy use patterns in the region. Additionally, a number of person-level characteristics are also controlled to ensure the representativeness of the synthetic population. The variables that are controlled in the population synthesis process are as follows:

- household level variables: income, household size, heat fuel type, tenure, structure type, number of rooms, household type, number of vehicles.
- person level variables: gender, age of householder, person race, employment status.

The TAZ level residential energy consumption is then calculated by aggregating the consumption of each synthesized household by TAZ.

5. Results

Figure 1 illustrates the TAZ level energy consumption per capita for different sources of energy. The results are displayed using the quantile classification method, with the lightest yellow representing the lowest 20% TAZs and dark brown representing the highest 20% TAZs in energy consumption per capita. The results for electricity energy consumption suggest each resident consumes approximately 19.43 million BTU (or 20.5 Billion Joules) 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 (20.6 Billion Joules) per year (U.S. Energy Information Administration, 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 (11.9 Billion Joules) per year, which is close to Georgia's consumption per capita (12.6 million BTU/13.3 Billion Joules per year). Natural gas consumption decreases in peripheral areas, where the heating fuel is primarily electricity or other fuel sources. Residents in peripheral areas tend to consume more energy generated by other fuel types. Overall, the energy consumption portfolio for a typical resident in the 10-county metro area is comprised of 69.2% electricity, 28.6% natural gas, and 2.2% other fuels. The decomposition of consumption by energy source is consistent with statistics at the Georgia statewide level (U.S. Department of Energy, 2016).

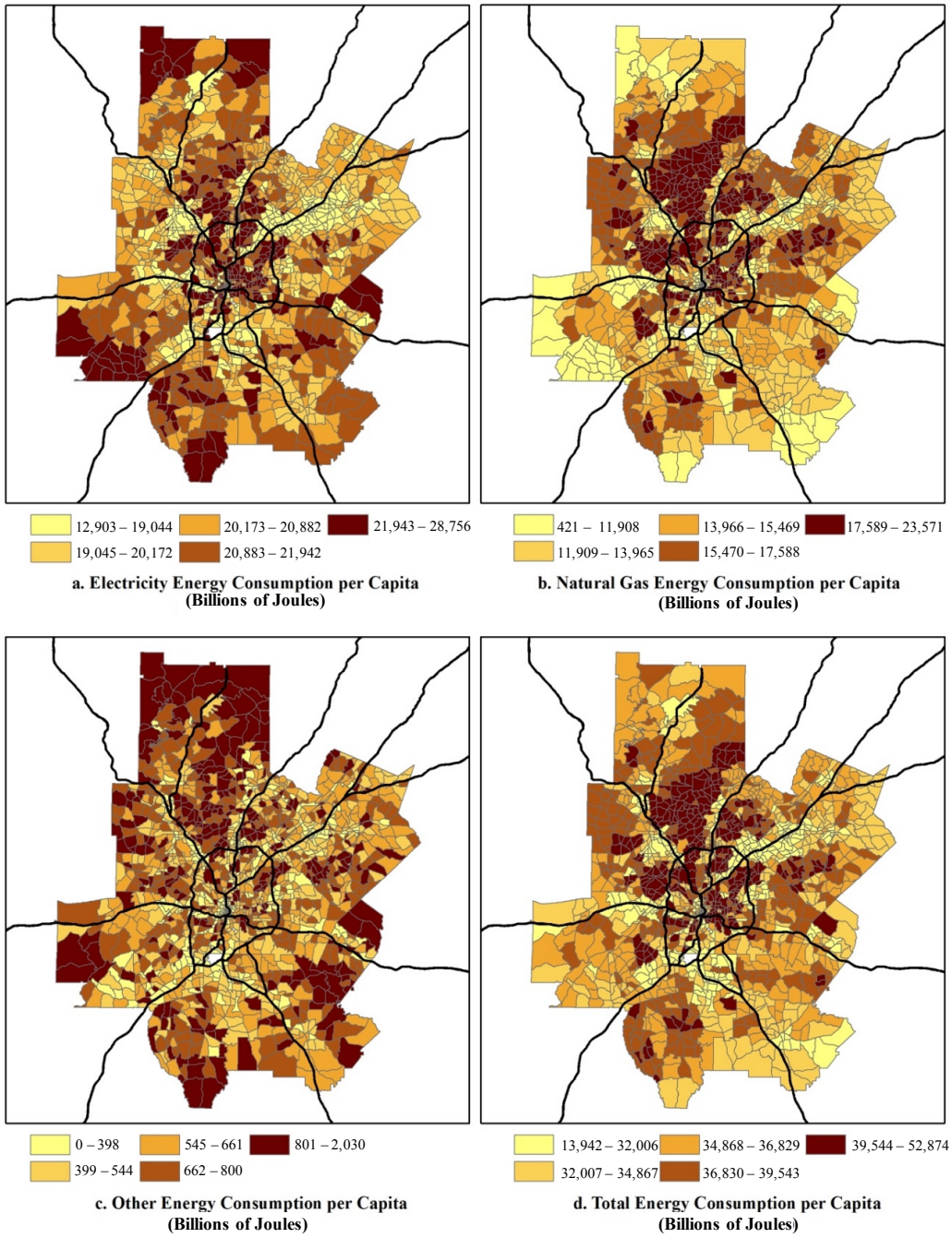

Figure 1: Synthesized TAZ level Annual Energy Consumption per Capita

The estimates of electricity and natural gas energy consumption are validated using the observed 2009 Atlanta energy consumption at the Zip Code level. The 2009 annual electricity consumption (in KWh) is provided by Georgia Power. The natural gas consumption (in Therms) is obtained from Atlanta Gas Light. The electricity consumption and natural gas consumption data are available for 31 and 42 Zip Codes correspondingly. There are 19 Zip Codes in the City of Atlanta. In the final analysis, Zip Codes whose centroids are outside of the city boundary, where the data providers are not the dominant energy suppliers, are excluded. After mapping TAZs to Zip Codes, estimates of energy consumption were aggregated to the zip code level (for

Zip Codes in the city). Overall, the correlations between model estimates and observed data are 0.908 and 0.927 for electricity and natural gas consumption respectively. The high correlations constitute a first indication of the validity of the proposed model framework. The average percent differences between the aggregated synthesized results and observed consumption data are respectively -13.6% and 5.5% for electricity and natural gas. It seems that the model underestimates residential energy consumption and overestimates natural gas consumption at the Zip code level, as shown in Figure 2 and 3. The differences for electricity consumption may be caused by the different definitions of residential sectors. The synthesized outputs do not include estimates for residents in group quarters, such as dorms, nursing homes, and institutions. However, the observed residential electricity data obtained from the utilities does include all types of residential units. For instance, there are the dorms of Savannah College of Art and Design located in zip code 30309, Georgia State University located in zip code 30303, and several boarding schools, such as the Westminster Schools and Atlanta Girls School, in zip code 30327, rendering the synthesized outputs are smaller than the observed electricity consumption. The difference in the natural gas consumption may be because Atlanta Gas Light is not the only gas supplier in some parts of the city, although it serves a vast majority of households. For instance, the natural gas consumption is overestimated in Zip code 30327, as shown in Figure 3, where households can choose between Atlanta Gas Light and Gas South. Additionally, the aggregation of the total population at the Zip code level may also lead to some errors, due to the discrepancies between Zip code, TAZ, and census block boundaries (TAZs are not perfectly nested within Zip Codes).

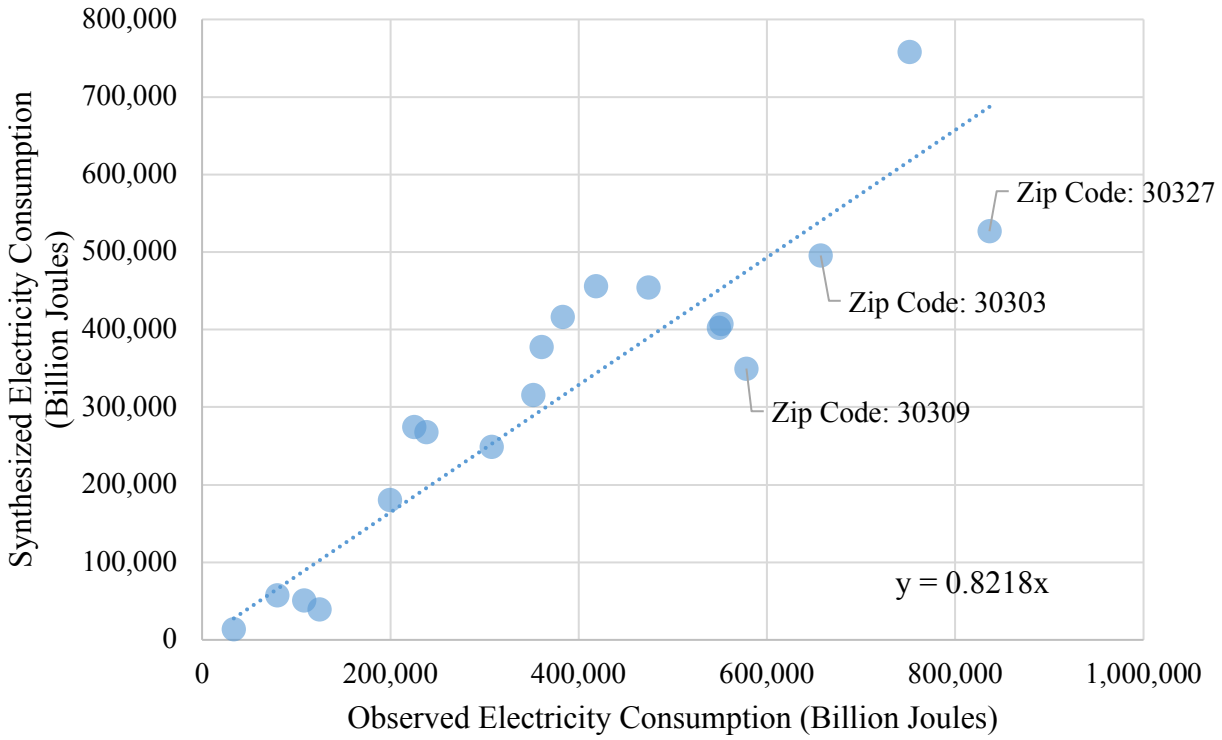

Figure 2: Electricity Synthesized Consumption vs. Observed Consumption by Zip Code

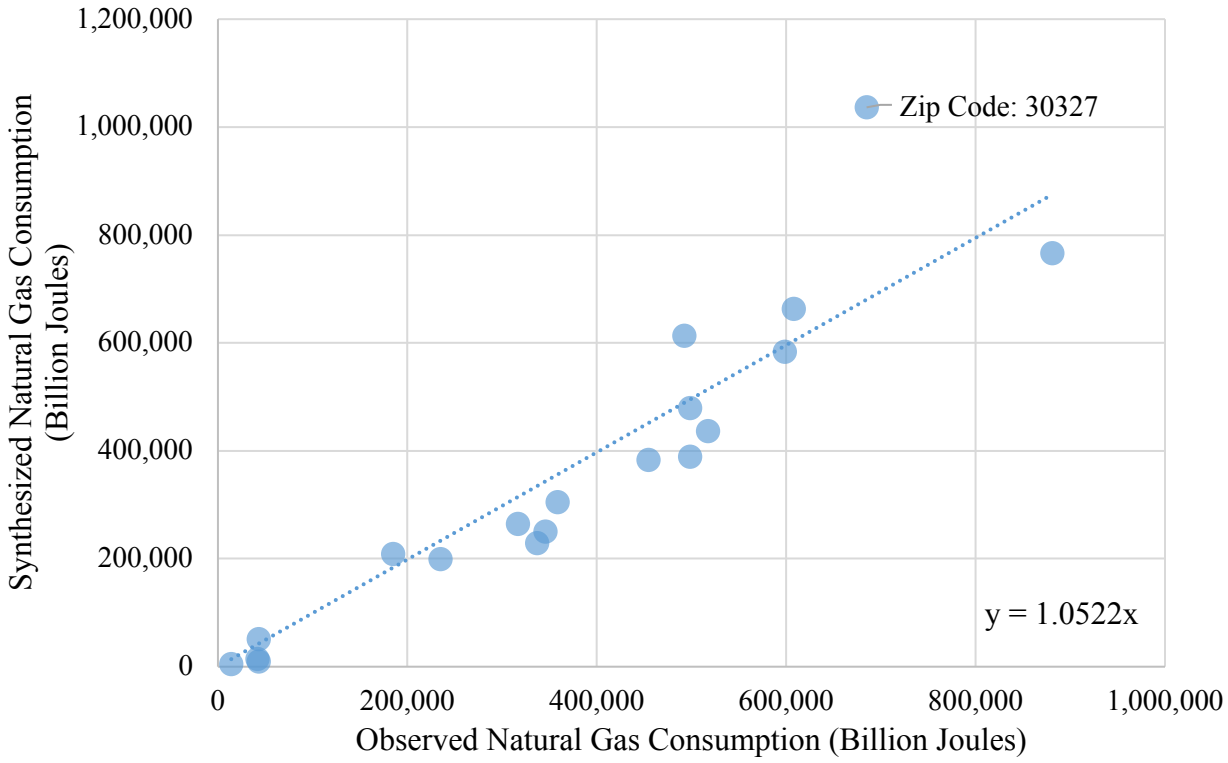

Figure 3: Natural Gas Synthesized Consumption vs. Observed Consumption by Zip Code

6. Conclusions

In this study, a residential energy consumption model system, which can be potentially applied to any major U.S. city, is developed. Most bottom-up energy estimation models are data intensive and therefore challenging to develop due to the difficulty in obtaining micro-sample data and regional housing inventories. To address these limitations, a bottom-up approach for energy consumption was developed in this study. This approach comprises a number of steps including statistical data matching, machine learning and household/population synthesis. The proposed model system requires data from publicly available sources including the RECS, PUMS, and ACS. The model outputs a set of synthesized households (representing the population of a region) with appended energy consumption by source.

The model was applied to the 10-county Atlanta metropolitan area with 1,593 TAZs to demonstrate its potential. The estimated results of residential electricity, natural gas, and other energy consumption per capita are 19.17, 14.05, and 0.59 million BTU per year, respectively. The results for Atlanta are consistent with statistics for the State of Georgia (U.S. Energy Information Administration, 2009). Additionally, model predictions were also cross-compared with actual electricity and natural gas utility information available for 21 zip codes. The correlations between our model results and observed consumption are 0.908 and 0.927 for electricity and natural gas use respectively. The results also show that electricity consumption is underestimated by 13.6%, while gas consumption is overestimated by 5.5%. The differences are quite modest and can be explained by the scope of the residential dataset captured in RECS, and the mismatch in geographic unit boundaries in the various datasets.

Ongoing energy footprint modeling efforts include integrating this model with transportation models (hence the use of TAZ as the geographic unit of interest), commercial buildings energy models, and life cycle energy consumption models to obtain a comprehensive energy footprint of neighborhoods and local zones within urban areas. This model only estimates residential energy consumption, but the machine learning approach can be applied to other energy sectors. The household's transportation energy consumption and embodied energy (i.e., the energy embodied in the construction and production of various products used by households) are also critical parts of the residential energy footprint, but not included within the scope of this paper.

Another area that merits future exploration is the analysis of the impact of alternative policy scenarios and land use forms on urban energy consumption. Our model can inform the energy consumption outcomes of alternative policies and urban forms. As cities continue to absorb growing populations and economic activity, invest in infrastructure, and implement zoning requirements, building codes, road pricing, and an array of other regulations and incentives, urban form will evolve; and so too will energy consumption patterns. As our empirical analysis of Atlanta demonstrates, urban form, socio-economic and demographic attributes, and building stock characteristics affect energy footprint and greenhouse gas emissions of metropolitan areas, providing decision makers a number of policy levers that can be exercised to foster more sustainable futures and resilient communities.

Acknowledgements

The authors are grateful to the Strategic Energy Institute at the Georgia Institute of Technology for providing partial funding to support this work. The authors are also grateful for the zip code level electricity and natural gas consumption data provided by Georgia Power and Atlanta Gas Light, cleaned and organized by Professor Valerie Thomas at Georgia Institute of Technology.

Appendix A Shared Variables in RECS and PUMS data

RECS 2009				PUMS 2009			
Variable	**Variable Descrip.**	**Response Codes and Labels**		**Variable**	**Variable Descrip.**	**Response Codes and Labels**	
TYPEHUQ	Unit structure	1 2 3 4 5	Mobile Home Single-Family Detached Single-Family Attached Apartment with 2-4 Units Apartment with 5+ Units	BLD	Unit structure	1 2 3 4 - 5 5 - 9	Mobile Home Single-Family Detached Single-Family Attached Apartment with 2-4 Units Apartment with 5+ Units
KOWNRENT	Tenure	1 2 3	Owned Rented Occupied without rent	TEN	Tenure	1-2 3 4	Owned with mortgage or clear Rented Occupied without rent
YEARMADE	Year housing unit was built	**C***	Year housing unit was built	YBL	When structure first built	01 02 03 04 05 06 07 08 09 10 11 12 13	1939 or earlier 1940 - 1949 1950 - 1959 1960 - 1969 1970 - 1979 1980 - 1989 1990 - 1999 2000 - 2004 2005 2006 2007 2008 2009
OCCUPYYRANGE	Year range when household moved in	1 2 3 4 5 6 7 8	Before 1950 1950 to 1959 1960 to 1969 1970 to 1979 1980 to 1989 1990 to 1999 2000 to 2004 2005 to 2009	MV	When moved into this house or apartment	1 2 3 4 5 6 7	12 months or less 13 - 23 months 2 - 4 years 5 - 9 years 10 - 19 years 20 - 29 years 30 years or more
BEDROOMS	Number of bedrooms	**C** -2	Bedrooms Not Applicable	BDSP	Number of bedrooms	**C**	Bedrooms
TOTROOMS	Number of Rooms	**C**	Rooms	RMSP	Number of Rooms	**C**	Rooms
FUELHEAT	Main space heating fuel	1 2 3 4 5 7 8 9 21 -2	Natural Gas Propane/LPG Fuel Oil Kerosene Electricity Wood Solar District Steam Other Fuel Not Applicable	HFL	House heating fuel	1 2 3 4 5 6 7 8 9	Utility Gas Bottled, tank, or LP gas electricity fuel oil, kerosene, etc. coal or coke wood solar energy other fuel no fuel used
NHSLDMEM	Number of members	**C**	Number of members	NP	Number of person	00 **C**	vacant number of person record
MONEYPY	Household income	01 02 – 22 23 24	Less than $2,500 $2,500 to $99,999 (with $5000 increases) $100,000 to $119,999 $120,000 or More	HINCP	Household income	00 -19998 **C**	No income Loss of 19,998 or more total income
DOLLAREL	Annual electricity cost	**C**	Dollars	ELEP	Monthly electricity cost	01 02 **C**	Included in rent or in condo fee no charge or electricity not used electricity bill paid (Dollars)
DOLLARNG	Monthly natural gas cost	**C**	Dollars	GASP	Monthly natural gas cost	01 02 03 **C**	Included in rent or in condo fee included in electricity no charge or gas not used gas bill paid (Dollars)
TOTALDOL	Annual total fuel costs	**C**	Dollars	FULP	Annual other fuel costs	01 02 **C**	Included in rent or in condo fee no charge or fuels are not used these fuel bill paid (Dollars)

*C: Continuous Variable

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.

Bianco, V., Scarpa, F., & Tagliafico, L. A. (2014). Analysis and future outlook of natural gas consumption in the Italian residential sector. *Energy Conversion and Management*, *87*, 754–764.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Brown, M. A., & Cox, M. (2015). Progress in energy and carbon management in large US metropolitan areas. *Energy Procedia*, *75*, 2957–2962.

Brown, M. A., Southworth, F., & Sarzynski, A. (2009). The geography of metropolitan carbon footprints. *Policy and Society*, *27*(4), 285–304.

Brown, M. A., & Wang, Y. (2015). *Green savings: how policies and markets drive energy efficiency: how policies and markets drive energy efficiency*. ABC-CLIO.

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.

Collins, M., Schapire, R. E., & Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, *48*(1–3), 253–285.

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.

D'Orazio, M. (2016). Statistical Matching and Imputation of Survey Data with StatMatch.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378.

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

Garikapati, V. M., You, D., Zhang, W., Pendyala, R. M., Guhathakurta, S., Brown, M. A., & Dilkina, B. (2017). Estimating Household Travel Energy Consumption in Conjunction with a Travel Demand Forecasting Model. *Transportation Research Record, Journal of the Transportation Research Board*.

Ghosh, S., & Kanjilal, K. (2014). Long-term equilibrium relationship between urbanization, energy consumption and economic activity: empirical evidence from India. *Energy*, *66*, 324–331.

Guhathakurta, S., & Williams, E. (2015). Impact of urban form on energy use in central city and suburban neighborhoods: lessons from the phoenix metropolitan region. *Energy Procedia*, *75*, 2928–2933.

Hargreaves, A., Cheng, V., Deshmukh, S., Leach, M., & Steemers, K. (2017). Forecasting how residential urban form affects the regional carbon savings and costs of retrofitting and

decentralized energy supply. *Applied Energy*, *186*, 549–561.
Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.
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.
Kavousian, A., Rajagopal, R., & Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, *55*, 184–194.
Konduri, K. C., You, D., Garikapati, V. M., & Pendyala, R. M. (2016). Application of an Enhanced Population Synthesis Model that Accommodates Controls at Multiple Geographic Resolutions. *Transportation Research Record, Journal of the Transportation Research Board*, *2563*, 40–50.
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.
Lopes, M. A., Antunes, C. H., & Martins, N. (2015). Towards more effective behavioural energy policy: An integrative modelling approach to residential energy consumption in Europe. *Energy Research & Social Science*, *7*, 84–98.
Nichols, B. G., & Kockelman, K. M. (2014). Life-cycle energy implications of different residential settings: recognizing buildings, travel, and public infrastructure. *Energy Policy*, *68*, 232–242.
Norman, J., MacLean, H. L., & Kennedy, C. A. (2006). Comparing high and low residential density: life-cycle analysis of energy use and greenhouse gas emissions. *Journal of Urban Planning and Development*, *132*(1), 10–21.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.
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.
Rässler, S. (2012). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches* (Vol. 168). Springer Science & Business Media.
Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222.
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.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
Tso, G. K., & Guan, J. (2014). A multilevel regression approach to understand effects of environment indicators and household features on residential energy consumption. *Energy*, *66*, 722–731.

U.S. Department of Energy. (2016). *Georgia Residential Energy Consumption*. Retrieved from https://apps1.eere.energy.gov/states/residential.html?state=GA

U.S. Energy Information Administration. (2009). *Household Energy Use in Georgia*. Retrieved from https://www.eia.gov/consumption/residential/reports/2009/state_briefs/pdf/ga.pdf

U.S. Energy Information Administration. (2016). *Energy Consumption by Sectors*. Retrieved from https://www.eia.gov/totalenergy/data/monthly/pdf/sec2_3.pdf.

Wang, S., Fang, C., Guan, X., Pang, B., & Ma, H. (2014). Urbanisation, energy consumption, and carbon dioxide emissions in China: a panel data analysis of China's provinces. *Applied Energy*, *136*, 738–749.

Yekang Ko. (2013). Urban Form and Residential Energy Use: A Review of Design Principles and Research Findings. *Journal of Planning Literature*, *28*(4), 327–351. https://doi.org/10.1177/0885412213491499

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

Zhang, W., Guhathakurta, S., & Ross, C. (2016). Trends in Automobile Energy use and GHG Emissions in Suburban and Inner City Neighborhoods: Lessons from Metropolitan Phoenix, USA. *Energy Procedia*, *88*, 82–87.

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