A Hybrid Dasymetric and Machine Learning Approach to High-Resolution Residential Electricity Consumption Modeling

April Morton, Nicholas Nagle, Jesse Piburn, Robert N. Stewart, Ryan Mcmanamay

Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831
Telephone: +1-865-576-9318
Email: {mortonam, naglenn, piburnjo, stewartrn, mcmanamayra}@ornl.gov

Abstract

As urban areas continue to grow and evolve in a world of increasing environmental awareness, the need for detailed information regarding residential energy consumption patterns has become increasingly important. Though current modeling efforts mark significant progress in the effort to better understand the spatial distribution of energy consumption, the majority of techniques are highly dependent on region-specific data sources and often require building or dwelling-level details that are not publicly available for many regions in the U.S. Furthermore, many existing methods do not account for errors in input data sources and may not accurately reflect inherent uncertainties in model outputs. We propose an alternative and more general hybrid approach to high-resolution residential electricity consumption modeling by merging a dasymetric model with a complementary machine learning algorithm. The method's flexible data requirement and statistical framework ensure that the model is both applicable to a wide range of regions and considers errors in input data sources.

Keywords: Energy Modeling, Dasymetric Modeling, Machine Learning.

1. Introduction

As urban areas continue to grow and evolve in a world of increasing environmental awareness, the need for detailed information regarding energy consumption patterns has become increasingly important. Since the residential sector alone accounts for approximately 30% of all energy consumption worldwide (Swan and Urgursal 2009), a detailed spatial understanding of residential energy consumption in particular is important for supporting efforts to promote conservation, technology implementation and other necessary changes in the energy infrastructure.

Since most open-access energy data is coarse and prevents the finer resolution spatial analyses needed to make meaningful decisions, several authors have proposed a variety of statistical and engineering-based techniques for estimating residential energy use at a more detailed level (Swan and Urgursal 2009). Though these methods mark significant progress in the energy modeling field, the majority are highly-dependent on region-specific data sources and often require building or dwelling-level details that are not publicly available for many regions in the U.S. Furthermore, many of these methods do not account for errors in input data sources and may not accurately reflect inherent uncertainties in model outputs.

In light of these limitations, we propose an alternative and more general hybrid approach to high-resolution residential electricity consumption modeling by merging the dasymetric model proposed by Nagle et al. (2014) with a complementary machine learning algorithm. Rather than basing the model off of sparsely available high-resolution data sources, we choose to disaggregate publicly available datasets into higher resolution target regions. The flexible data requirement, along with the model's statistical framework, ensures that the model is both applicable to a wide range of regions and considers errors in input data sources.

2. Methodology

To create a flexible high-resolution energy modeling framework, we borrow techniques from the fields of dasymetric modeling and machine learning. The goal of dasymetric modeling is to produce high-resolution estimates by disaggregating total populations into smaller target regions through ancillary data layers and an appropriate model (Slocum el al. 2009). One of the primary goals of machine learning is to make useful predictions based on algorithms developed from existing data (Breiman 2001). We use the dasymetric approach proposed by Nagle et al. (2014) to disaggregate a weighted sample of surveyed households into smaller target geographies and then use a complementary learning algorithm to estimate electricity consumption for each of the disaggregated households. We then sum over the estimated consumption values for households placed within specific target regions to produce aggregate estimates. In the following section we discuss the hybrid dasymetric and machine learning approach to residential electricity consumption modeling in greater detail.

2.1 The Hybrid Dasymetric and Machine Learning Approach to High-Resolution Residential Electricity Consumption Modeling

Suppose we have access to a source sample containing n of N households from region s partitioned into $t \in \{1, \dots, T\}$ target regions and assume p_{it} is the unknown probability that household i is in target region t. In addition, assume each survey response contains a variable or vector of variable values c_i which can be used by some learning function $f(c_i)$ to estimate household electricity consumption.

Our first goal is to estimate the expected number of households $w_{it} = Np_{it}$ that are like household i in target region t by using complementary ancillary data and a dasymetric model. Suppose we are given this ancillary information in the form of total households with specific characteristics for nested geographies within s. More specifically, let $\widehat{pop}_{ka}$ represent an uncertain estimate of the number of housing units with characteristic k in sub-region a (i.e. the number of 5-bedroom housing units in a specific tract in Tennessee) where e_{ka} is the positive or negative error between the estimated and true housing unit count and σ_{ka}^2 is the variance of the error. Given the above constraints, as well as prior probabilities q_{it} for all p_{it}, we determine the number of households $w_{it} = Np_{it}$ that are like household i in target region t, as well as each of the errors e_{ka}, by solving the optimization problem

$$\min n \sum_{it} p_{it} \log \frac{p_{it}}{q_{it}} + \sum_{ka} \frac{e_{ka}^2}{2\sigma_{ka}^2} \tag{1}$$

subject to the relaxed pycnophylactic constraints

$$\sum_{i}\sum_{t\subseteq a} w_{it}\cdot I_k(w_{it}) = \widehat{pop}_{ka} + \ e_{ka} \text{ for each constraint } k \text{ and sub-region } a \tag{2}$$

where

$$I_k(w_{it}) = \begin{cases} 1 \text{ If household } i \text{ has characteristic } k \\ 0 \text{ Otherwise} \end{cases} \quad . \tag{3}$$

The term $\sum_{it} p_{it} \log\frac{p_{it}}{q_{it}}$ in the objective function is equal to the Kullback-Leibler divergence of q from p. Thus, one way of interpreting our objective function is both choosing the p that minimizes the information lost when using the prior distribution q to approximate p and choosing the error terms that minimize the sum of the ratio of squared error terms over variance terms.

Once the above estimates w_{it} have been determined, the final aggregate electricity consumption for target region t can be estimated through the function

$$C_t = \sum_{i} w_{it}\cdot f(c_i). \tag{4}$$

3. Application and Results

To illustrate the utility of the proposed hybrid model we estimate the aggregate electricity consumption at the block group level for Anderson County, Union County, and Knox County of the Knoxville Metropolitan Statistical Area.

3.1 Datasets

We obtain the source population from the 2008-2012 household-level Public Use Microdata Sample (PUMS) of the American Community Survey (ACS) (U.S. Census 2012), which includes detailed demographic and household characteristics, as well as a variable denoting whether the average monthly electricity cost is provided and if so its value in dollars per household, for a 5 percent sample of households chosen from course geographic units called PUMAs (U.S. Census 2012). To determine the constraints and variances we use the 2008-2012 ACS summary tables (U.S. Census 2012), which contain both tract and block group level average totals and their corresponding 90% margins of error (moes) (U.S. Census 2009).

The weights w_{iu}^{s} provided for each household by the PUMS survey represent the number of households that are like househould i in PUMA u. We assume each unique household has the same probability of belonging to each target region and thus let $q_{it} = \frac{w_{iu}^{s}}{T\cdot\sum_{r=1}^{n} w_{ru}^{s}}$. We select the tract and block group-level ACS summary table totals summarized in table 1 as our constraints $\widehat{pop}_{ka}$ and their corresponding 90% moes as our error variances σ_{ka}^{2}. When the variable denoting the average monthly electricity cost is provided we use it to estimate c_i and then compute the average monthly consumption through the function $f(c_i) = c_i/p_i$, where $p_i = 0.097$ is the average cost per kWh reported by the Knoxville Utilities Board (Knoxville Utilities Board 2012). When the average cost is not provided we estimate c_i using Breiman's Random Forest Regression trained with the same constraint variables and again use the function $f(c_i) = c_i/p_i$ to estimate the average monthly consumption in kWh.

Constraint	Tract	Block Group
0 person households	X	X
1 person households	X	
2 person households	X	
3 person households	X	
4 or more person households	X	
0 bedroom households	X	X
1 bedroom households	X	X
2 bedroom households	X	X
3 bedroom households	X	X
4 bedroom households	X	X
5 or more bedroom households	X	X
Households built 2010 or later	X`	X
Households built 2000 to 2009	X	X
Households built 1990 to 1999	X	X
Households built 1980 to 1989	X	X
Households built 1970 to 1979	X	X
Households built 1960 to 1969	X	X
Households built 1950 to 1959	X	X
Households built 1940 to 1949	X	X
Households 1939 or earlier	X	X
Housing units	X	X

Table 1. Tract and Block-Level Constraints

3.2 Results and Discussion

Fig. 1 shows the normalized estimated average block group-level electricity distribution in kWh per m^2 for Anderson County, Union County, and Knox County. As expected, the block groups closer to the Downtown Knoxville area and University of Knoxville, Tennessee have a much higher normalized average residential electricity consumption than the more rural areas lying outside the major cities. In addition, we see that most of the uninhabited areas, such as Blue Ridge State Park and the forested areas to the west of Rocky Top have very low normalized average residential electricity consumption estimates.

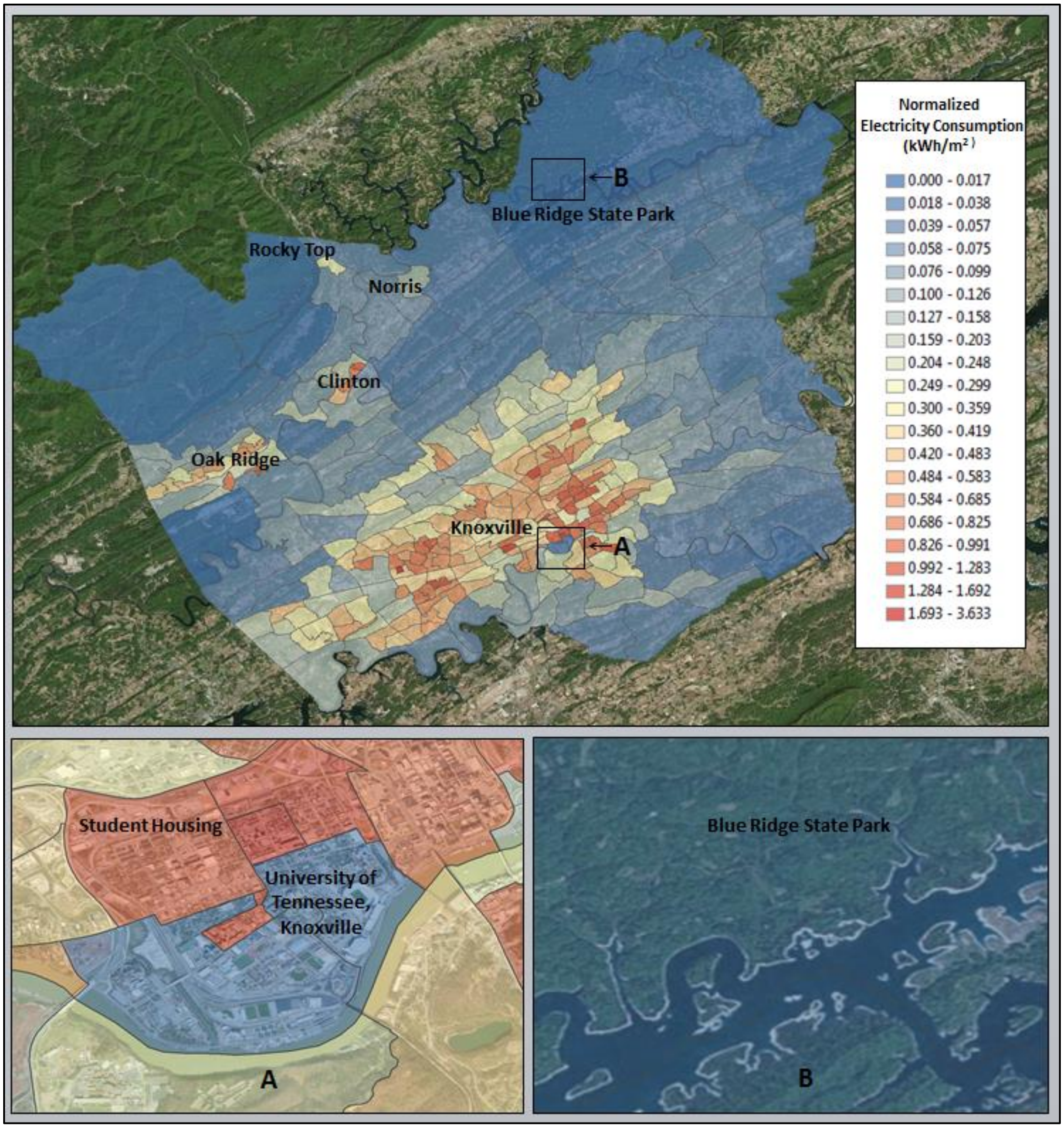

Figure 1. Normalized Block Group-Level Residential Electricity Consumption Estimates (kWh/m^2) for Anderson, Union, and Knox Counties

4. Conclusion

In this paper we presented a novel hybrid dasymetric and machine learning approach to high-resolution residential electricity consumption modeling and demonstrated the utility of the method by using it to estimate and analyse aggregate block group level residential consumption within a growing urban area. The model overcomes existing limitations by only requiring commonly available data and providing a well-defined method for handling uncertain input data sources. Furthermore, the dasymetric framework provides new opportunities for other scientific areas that would benefit from finer resolution spatial analyses using existing open-access information.

5. Acknowledgements

This manuscript has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

6. References

Breiman, L, 2001, Random forests, *Machine learning* 45(1): 5–32.

Knoxville Utilities Board , 2012, Building for the Future. [online] Knoxville, p.8. Available at: http://www.kub.org [Accessed 6 Feb. 2015].

Nagle, NN, Buttenfield, BP, Leyk, S and Spielman, S, 2014. Dasymetric modeling and uncertainty. *Annals of the Association of American Geographers*, 104(1):80–95.

Slocum, TA, McMaster, RB, Kessler , FC, and Howard, HH, 2009, *Thematic cartography and geovisualization*, Upper Saddle River, NJ: Pearson Prentice Hall.

Swan, L. G. and 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), pp. 1819–1835.

U.S. Census Bureau, 2009, *A compass for understanding and using American Community Survey data: What researchers need to know*, Washington, DC: U.S. Census Bureau.

U.S. Census Bureau, 2012, 2008 – 2012, American Community Survey Microdata. [online] Available at: http://factfinder2.census.gov [Accessed 6 Jan. 2015].

U.S. Census Bureau, 2012, 2008 – 2012, American Community Survey Summary Tables. (2012). [online] Available at: http://factfinder2.census.gov [Accessed 6 Jan. 2015].