1
2
3
4
5
6
7 **Defining Geographical Boundaries with Social and Technical Variables to Improve**
8 **Urban Energy Assessments**
9

10 Janet L. Reyna^{1*}, 660 S. College Ave. Tempe, AZ 85281 USA, janet.reyna@asu.edu
11

12 Mikhail V. Chester¹, 660 S. College Ave. Tempe, AZ 85281 USA, mchester@asu.edu

13 Sergio J. Rey², 976 S. Forest Mall, Tempe, AZ 85281 USA, srey@asu.edu

- 14 1. Civil, Environmental, and Sustainable Engineering, Arizona State University
15 2. School of Geographical Sciences and Urban Planning, Arizona State University
16 * Corresponding Author: janet.reyna@asu.edu, +1-918-605-1874
17

24   **Abstract**

25   Within residential electricity consumption there exists significant variability from
26   home-to-home due to the differences in building thermal properties, appliances, and
27   inhabitants. Electricity analyses at sub-city scales using predefined geographies, such
28   census tracts, might artificially split areas with homogenous characteristics leading to
29   analyses that don't effectively contrast the drivers of energy use. The objective of
30   this study is to use the spatial relationships between demographics, building types,
31   and electricity consumption to form new geographies with less variability for use in
32   residential energy assessment. Using Los Angeles and New York City as case studies,
33   differences in variability of energy use within predefined geographies (e.g., census
34   tract) are compared to geographies defined by clustering on socio-technical
35   characteristics. Socio-technical clustering, regardless of the chosen subset of
36   variables, reduces the variability over pre-defined geopolitical boundaries with high
37   statistical significance ($p \ll 0.0001$). By clustering for energy analysis, intra-regional
38   variability is reduced by 13% in Los Angeles and 29% in New York, improving
39   opportunities for prediction and forecasting. This is the first study to examine the role
40   of spatial boundaries in urban energy assessment. The creation of socio-technical
41   geographies for electricity assessment creates opportunities for improving predictions
42   and forecasts for future sub- and cross-city energy studies.

43

44 Nomenclature

CBG	U.S. Census Block Groups
CT	U.S. Census Tracts
NYC	New York City
LA	Los Angeles
LADWP	Los Angeles Department of Water and Power
ACS	American Community Survey
RASS	Residential Appliance Saturation Survey
RECS	Residential Energy Consumption Survey
HoH	Head of Household
HH	Household

45

46

1 Introduction

Residential building energy use is a major driver of urban energy use, yet there remains relatively poor information at neighborhood scales across cities of the drivers of this energy consumption. In 2014, residential buildings consumed 21.7% of total energy use in the U.S. [1], nearly 60% of which was electricity [2]. Understanding how and where homes within cities use energy will become even more critical as electricity demand in U.S. homes has increased 34% over the last three decades despite efficiency improvements [3]. Improving understanding of how energy is used in homes across cities will require analysis of large geographical regions with sufficient detail to support informed decision-making. This article assesses whether the use of newly defined sub-city geographies for building energy use based on social and technical variables yields more meaningful predictive capacity than predefined (and often geopolitical, such as census tracts) regions.

1.1 Energy and Spatial-Scale

To date there are few rigorous city-wide assessments of building energy use. This research defines four characteristics of a rigorous study of building energy use: 1) wide geographical coverage (e.g. an entire city), 2) a large sample size or universal coverage of the area, 3) spatial detail at an intermediate scale so that different areas of energy consumption in different areas of the city can be identified, and 4) validation with large-scale reporting of actual energy use. Few studies meet all four of these criteria. Many building electricity analyses focus on the individual building level. For example, Chan investigates the optimal height for installing balconies on residential buildings in order to maximize the solar shading benefits [4], and Pan et

al. examine green building design options to improve two office buildings in Shanghai, China [5]. However, extrapolating from a single building to predict a larger area may fail to capture the variability between different types of buildings and different habits of users. Another group of studies investigate city-level or national patterns with minimal spatial resolution. Tso and Guan perform a multilevel regression on a residential energy survey that maintains regional (i.e. 1-4 U.S. states) resolution, but without sub-urban detail [6]. Larivière and Lafrance develop a statistical model that relates city-wide variables such as density, weather, and demographics to electricity use in Canada [7]. Howard et al. 2012 model energy use across New York City at fine spatial resolutions by applying end-use intensities to building type and floor area [8]. Although their data are modeled and not metered, this type of detailed geospatial analysis could be an important framework for capturing macro patterns while retaining building-level detail. Anderson et al. find that building-level and urban-level energy studies tend to be confined within their respective scales with little overlap, but argue that an integrated framework is necessary to achieve policy goals within the building sector [9]. In building energy assessment, working at a meso-scale is necessary to capture both wide geographical coverage and spatial detail.

City-wide energy analyses need meso-scale spatial detail to understand how consumption varies among different users and areas of a city, but efforts modeling at this scale do not meet the four criteria outlined in the previous section. In this article, "meso" is defined as smaller than urban or census tract, but larger than an individual household or city block. It can be thought of as a neighborhood, or group of neighborhoods with similar characteristics. Furthermore, the choice of the meso-

93 geography needs to be based upon characteristics relevant to energy consumption.
94 Building energy use is spatially correlated due to micro-climates from urban form [10]
95 and also through relationships in building design, year of construction, and type of
96 inhabitants. There have been some efforts to model urban energy-use at the sub-city
97 level, but nearly all of these efforts are simulation-based from a handful of building
98 types. For example, Shimoda et al. simulate energy consumption in Osaka City using
99 archetypal use patterns [11], and Jones et al. model residential energy used based on
100 characteristics of the building stock [12]. In a similar way, Stephan and Stephan study
101 the life cycle energy use of residential buildings in Lebanon based off of a single case
102 study [13]. Building energy studies to date that do include meso-levels of geography
103 tend to default to geopolitical boundaries such as the census tract, without any
104 justification for the choice of spatial boundaries. Caputo et al. analyze energy policy
105 strategies in the building sector at the neighborhood scale in Milan, Italy, but
106 aggregates the study to the Italian National Census Tract level [14]. Heiple and Sailor
107 develop a bottom up model for residential and commercial buildings in Houston,
108 Texas, and aggregate the parcel-level data into 1.33 kilometer grid cells [15]. There
109 are no other studies found that investigate the impact of meso-geographies on intra-
110 regional heterogeneity in building energy use. There is a need not only to study
111 energy use at the meso-scale, but also to develop methods for aggregating data to
112 these intermediate geographies. To fill this gap, the objective of this study is to use
113 the spatial relationships between demographics, building types, and electricity
114 consumption to form new geographies with less variability for use in residential
115 energy assessment. This is the first study to quantify the impact of spatial boundary

selection on energy use homogeneity. In this article, the terms “homogeneity” and “heterogeneity” are used to discuss the similarity or dissimilarity of data points within a region (e.g. demographic data or energy consumption).

1.2 Redefining Energy Assessment Geographies

In this study, meso-level spatial boundaries are created for energy assessment based upon the social and technical characteristics of residential units. Social characteristics are defined as the demographic profile of the households, and the technical characteristics are defined as the building shell properties along with appliance prevalence and use. This approach has the advantage of capturing both the geographic and socio-technical definitions of “neighborhoods”, and energy analysis geographies should include not just what kinds of buildings are there, but also what kinds of people use those buildings. Using pre-defined geographies is potentially problematic as it might split areas of relatively homogeneous electricity use thereby making predictions less accurate. Literature on zone choice exists in the social sciences for identifying spatial correlations in demographic data or assessing boundary changes over time. For example, Rey and Sastré-Gutiérrez (2010) compare different regionalization schemes in assessing regional inequality in Mexico [18]. The definition of appropriate geographical boundaries is an integral component to large-scale energy analysis, but there is no identified literature discussing the significance of zone choice on energy assessments.

1.3 Case Studies

To explore the significance of geography definitions to the prediction of electricity use, two case studies are developed for the residential building stocks of two United

States cities: Los Angeles (LA), California, and New York City (NYC), New York. Residential stocks are chosen because the function of (and to some extent activities within) residential buildings is largely homogenous, unlike for commercial or industrial buildings.

For LA and NYC, geographies built using endogenous socio-technical characteristics are compared to geopolitical (i.e. census tract) geographies to examine which has more homogenous electricity consumption. The correlation between social and technical characteristics of homes and electricity consumption is tested, and as is whether the use of these characteristics in clustering can reduce the variance in electricity predictions. This hypothesis is tested in two cities with unique building stock, density, consumption profiles, and climates.

2 Methodology

Electricity, social, and technical data are prepared at the census block group (CBG) level, and a max-p clustering algorithm is used to explore the significance of geography in electricity analysis. It is hypothesized that clustering on social and technical variables using a max-p clustering algorithm will reduce heterogeneity of electricity use between residential buildings within each cluster over using pre-defined geographical boundaries. The clustering algorithm is used to aggregate CBGs into larger areas by creating spatially-contiguous areas with minimum difference within a set of pre-defined variables. The goal is to identify the clustering formulations which lead to the most homogeneous regions with respect to electricity consumption. For each city, there are nine sets of variables used for clustering:

- 161 1) Electricity use alone. This provides the most homogenous regions possible
162 within a given data set since electricity use is the variable of interest.
- 163 2) Technical variables on building characteristics and appliance prevalence.
- 164 3) Social variables on the demographics of the inhabitants
- 165 4) A combination of all social and technical variables
- 166 5) A subset of the 10 most significant social and technical variables
- 167 6) A subset of the 6 most significant social and technical variables
- 168 7) All social and technical variables with weights representing the significance of
169 each variable to total energy use
- 170 8) The subset of the 10 most significant social and technical variables with
171 weights representing the significance of each variable to total energy use
- 172 9) The subset of the 6 most significant social and technical variables with weights
173 representing the significance of each variable to total energy use
- 174 These clustering formulations inform the types of variable subsets that might be
175 significant to forming new energy analysis geographies for each city. Figure 1 gives an
176 overview of the methods.

Figure 1: Process overview. Visual summary of methodological steps.

2.1 Geospatial Electricity Demand

CBG electricity data is obtained for both cities, and census and appliance surveys are used to calculate social and technical variables for each CBG. Household-level electricity data are rarely available for large areas and are more frequently obtained at an intermediate geography such as CBG or zip-code. Likewise, social data are generally at a meso-geography, for example census tract, and not the household level. For this reason, CBG data is used to test the hypothesis, although similar principles could be used to cluster at the household level if data are available.

For the case studies, the study scope includes only electricity use and not natural gas consumption for several reasons. For the NYC case study, there are estimates of

both electricity and fuel (including natural gas and other heating fuels), but there is only electricity data for LA. The hypothesis is therefore tested on electricity data, although this methodology is likely as applicable for analyzing total home energy use. In the Appendix, there is a case study for NYC comparing how clustering results differ between electricity versus total home energy use.

For LA, total annual residential electricity consumption is obtained for fiscal year 2011-2012 from the LADWP service area, aggregated by CBG [19]. LADWP is the electricity provider for approximately 40% of the residential accounts in Los Angeles County, which is 2503 CBGs and 1,305,759 residential accounts. The LADWP service territory covers 1,157 km².

The NYC Mayor's Office of Long-Range Planning and Sustainability provides electricity data at the zip-code level for year 2009, but these data are not resolute enough to formulate questions on geography. Howard et al. 2012 developed a regression model for annual consumption using zip-code data which predicts electricity and fuel use at the tax-lot (a geography typically containing 1-2 buildings) by using building use code and year of construction as predictors [8]. The predictive results of this study generally fall within 20% of the actual data. For this study the residential subset from the Howard model is used to test the clustering hypotheses, covering all of NYC, 587 km².

For NYC, electricity and fuel use on each tax-lot is allocated between residential and non-residential purposes with the factors provided in Howard et al. 2012. The number of residential units on each tax lot is used as an approximation of the number of electricity accounts, although this might not always be the case (e.g.

an entire building is metered under a single account). In this study, the household is the basic unit of energy use, and residential units stand as a reasonable approximation for household count. Once residential electricity and fuel consumption is partitioned at the tax-lot level, these values are aggregated to the CBG. For NYC, data is included from all five boroughs, which encompasses 6497 CBGs and approximately 3.8 million electricity accounts (Figure 2).

Figure 2: Household electricity use in LA and NYC. Annual household electricity use in LA and NYC by CBG. Overall, NYC has higher electricity consumption and more variance between areas.

2.2 Socio-Technical Data

For LA and NYC, social and technical profiles are compiled for every CBG in each city. The American Community Survey (ACS), conducted annually by the U.S. Census Bureau, provides a rich array of demographic variables, including many which previous studies showed to be influential to electricity consumption [20]. ACS provides these data at the CBG level and combines five years of estimates. For this study, the 2008-2012 ACS data are used in both LA and NYC. Additionally, in both cities, governmental agencies have residential energy consumption surveys, including

information on building stock characteristics and appliance prevalence. However, these surveys lack information on the spatial distribution of these technical characteristics.

Technical properties of buildings are assumed to be correlated to 1) the building classification (single family detached, multifamily high-rise, etc.) and 2) the year of construction, and a prototyping approach is employed to map survey data to geographies. For LA, a subset of data for the LADWP service area is pulled from the 2009 California Residential Appliance Saturation Survey (RASS) which is developed by the California Energy Commission[21]. RASS micro-data are household-level, and the LADWP subset includes approximately 2697 survey responses. In NYC, the 2009 Residential Energy Consumption Survey (RECS), conducted nation-wide by the U.S. Energy Information Administration (EIA), is used [22]. Both the RASS and RECS datasets provide household-level survey data on a variety of socio-technical characteristics relevant to electricity consumption as well as the quantities of electricity, natural gas, and other fuels used in the home.

Household-level data from RECS is filtered to obtain a subsample that is NYC-specific. In RECS as in RASS, the EIA provides household-level micro-data, but for most of the country these data are only given by geographic region (e.g. "West South Central Census Division" which includes Arkansas, Louisiana, Texas, and Oklahoma). Each time RECS is administered, the micro-data provided additional geographical resolution for several of the largest U.S. states such as Texas and California. This is the case for New York state as well. The RECS micro-data is further filtered to obtain only households in NYC by using two other criteria: climate zone and urban/rural

classification. Most of New York state falls into the "Very Cold/Cold" Building America Collapsed Climate Region [23], however the five counties of NYC and three surrounding suburban counties are "Mixed-Humid". Data from the three surrounding suburban counties are removed by including only urban data points. Thus the 475 household-level data points for NYC are obtained using the boundaries of the climate zones and urban/rural classification.

2.3 Variable Selection and Assignment

Nine social variables and twelve technical variables from the RASS, RECS, and ACS datasets are selected which are significant to residential electricity use based on previous studies (Table 1). RASS and RECS variables that are as similar as possible are selected to allow for an equivalent comparison between the two regions. Slight differences exist between the variable definition of the RASS and RECS datasets; these differences are discussed in the Appendix.

Table 1: Socio-technical variable subset for max-p clustering. The variables were selected based on their availability in the survey data and their significance in previous literature. The variable, name, description, and source are provided as well as the studies that have identified these characteristics as significant.

Variable Name	Type	Description	Literature
Age	Social	Head of household (HoH) age	[7,24-27]
Family Composition	Social	Household (HH) is a family unit	[25,24,28,29]
Gender	Social	HoH is female	[25,26]
HH Size	Social	Number of members in HH	[24-27,29-33]
Education Level	Social	HoH has only a HS degree or less	[29]
HH Income	Social	HH Income	[7,25,26,24,29,32,3031,33]
Ethnicity	Social	HoH is non-white	[25,24]
Rental Unit	Social	HH in rental unit	[24,27,29,33]
Employment	Social	HoH is unemployed	[24]

Air Conditioning	Technical	Unit has air conditioning	[24,30,31,33,34]
Insulation	Technical	HoH reports "adequate" insulation	[35,25,27,29,36,37]
Construction Year	Technical	Building year of construction	[25,26,33,38]
Windows	Technical	Unit has double-paned windows	[35,27,30,36,37]
Electric Heat	Technical	Unit has electric heat	[7,27,29,34]
Exterior Siding	Technical	Unit has wooden or vinyl siding	[35,37]
Natural Gas Use	Technical	Cubic feet/therms of natural gas use	[25]
Refrigerators	Technical	Number of refrigerators in unit	[24,27,31,34]
Computers	Technical	Number of computers in unit	[34]
Televisions	Technical	Number of TVs in unit	[27,31,34]
Pool	Technical	Unit has a pool	[29,31,27]
Floor Area	Technical	Unit size	[25,26,29,30,27]

272

273   Technical data are assigned to each CBG based upon the age and type of the
274   residential building stock within each area. From the micro-data available in the RASS
275   and RECS dataset, thirty-two prototypes are developed for LA and eighteen prototypes
276   for NYC. NYC has fewer prototypes because there are fewer sample points in the
277   dataset, and categories had to be collapsed to have a representative sample for each
278   prototype. The full list of these prototypes and their sample sizes is provided in the
279   Appendix. For every prototype category a technical profile of the characteristics
280   listed in Table 1 is calculated.

281   Every building in LA and every tax lot in NYC is assigned a prototype category
282   based upon the building use code and year of initial construction. In LA, the 2009
283   County tax assessor's database is used for building-level classification [39], and in NYC
284   the PLUTO database is used [40]. From these two data sources, the number of
285   buildings of each prototype category located in each CBG is known for the case study

regions. From these, technical profiles are assigned based upon the weighted average of the properties of the prototypes in each CBG. This gives a unique technical profile for every CBG in each city.

2.4 Clustering Formulation

The max- p regions clustering algorithm, a geospatial assessment approach that relies upon the properties of existing small regions to endogenously determine the number of clusters, is used in developing new geographies. This algorithm was developed specifically for defining geographical boundaries for quantitative analysis [41]. Applications of the algorithm to-date include assessing changes in neighborhood boundaries across time [42] and studying inter-regional equity [18]. The algorithm aggregates n geographic areas into the maximum number of regions p such that:

1. All regions are spatially connected
2. The value of a predefined attribute for each region must be greater than or equal to a predefined value
3. Each area A_i must be assigned to a single region R_k
4. Each region R_k must have at least one area A_i [41]

Setting an appropriate predefined threshold value (Item 2) is an important condition for the solutions of this algorithm. Other more traditional clustering approaches include setting the number of clusters a priori. For this approach, the threshold is set to provide a similar number of residential accounts as a census tract, but since this algorithm does not provide a deterministic solution, there is some variability in the clusters the algorithm forms. This variability is explored in the results and discussion. The predefined threshold value is the number of electricity

309 accounts; the minimum number of electricity customers in each solution is therefore
 310 specified.

311 2.4.1 Max-P Clustering Algorithm

312 A mathematical overview of the max-p clustering algorithm is provided, adapted from
 313 Duque et al. 2012 [41] for this specific problem:

314 *Step 1: Define Set of Areas for Los Angeles County (Census Block Groups)*

$$315 \quad A = \{A_1, A_2, \dots, A_n\}$$

$$316 \quad n = |A|$$

317 *Step 2: Define attributes of each census block group, and the set of attributes*

318 A_{iy} , denotes attribute y of area A_i where $y \in Y = \{1, 2, \dots, m\}$ with $m \geq 1$

319 *Step 3: Define the dissimilarity between census block groups*

320 $d: A \times A \rightarrow \mathbf{R}^+ \cup \{0\}$ is the dissimilarity based on Y such that $d_{ij} \equiv d(A_i, A_j)$ satisfies

$$321 \quad d_{ij} \geq 0, d_{ij} = d_{ji}, \text{ and } d_{ij} = 0 \text{ for } i, j = 1, 2, \dots, n$$

322 *Condition: All areas must be touching and form a single region*

323 $W = (V, E)$ denotes contiguity graph so that $v_i \in V$ corresponds to areas $A_i \in A$ and

324 edges $\{v_i, v_j\} \in E$ if and only if A_i and A_j share a boarder. W must be connected graph

325 *Step 4: Defining all feasible partitions*

326 $P_p = \{R_1, R_2, \dots, R_p\}$ partition of areas A into p regions with $1 \leq p \leq n$ such that:

327 $|R_k| > 0$ for $k = 1, 2, \dots, p$ (each area must contain at least one area A_i)

328 $R_k \cap R_{k'} = \emptyset$ for $k, k' = 1, 2, \dots, p \wedge k \neq k'$ (there is no overlap between areas)

329 $\cup_{k=1}^p R_k = A$ (adding all areas R_1 through R_p will give the full area A)

330 The sum of the attribute of interest, l_i for each area A_i within R_k must be greater than
 331 the defined threshold, and the threshold must be defined as to be less than the sum
 332 of l_i for all areas A_i

$$333 \quad \sum_{A_i \in R_k} l_i \geq \text{threshold} \begin{cases} \text{for } k = 1, 2, \dots, p, \text{ and} \\ \text{threshold} \in \mathbf{R}^+ \cup \{0\} | 0 \leq \text{threshold} \leq \sum_{A_i \in A} l_i \end{cases}$$

334 $W(R_k)$ is connected for $k = 1, 2, \dots, p$

335 Let Π denote the set of all feasible partitions of A

336 *Evaluation Criterion for a Feasible Partition $P_p \in \Pi$*

337 Heterogeneity of region k with $R_k \in P_p$ (summation of dissimilarity between all areas
 338 in R_k):

339

340

$$341 \quad h(R_k) = \sum_{ij: A_i, A_j \in R_k, i \leq j} d_{ij}$$

342 Total heterogeneity of partition $P_p \in \Pi$

$$343 \quad H(P_p) = \sum_{k=1}^p h(R_k)$$

344 Max-p-regions problem: maximizes the number of regions in the partition and
 345 maximizes the heterogeneity between the regions.

346 Determine $P_p^* \in \Pi$ such that $|P_p^*| = \max(|P_p|: P_p \in \Pi)$, and

347 $\nexists P_p \in \Pi: |P_p| = |P_p^*| \wedge H(P_p) < H(P_p^*)$

348 A summary of the variables included in the formulation are given below (Table 2).

349 Table 2: Variable Definitions

Variable	Definition
n	Number of areas in Set A
A	Set of areas A_1 through A_n
A_{iy}	Denotes attribute y in area A_i
l_i	Spatially extensive attribute of area A_i
d	Dissimilarity between the areas
P_p	A set of partitions of A into R
R_k	A grouping of areas A
Π	Set of all feasible partitions of A
$h(R_k)$	Heterogeneity of R_k
$H(P_p)$	Heterogeneity of a partition P_p

350

351 2.4.2 Algorithm Implementation

352 Clustering on electricity data alone to provides a lower bound on the
353 heterogeneity of electricity consumption that naturally exists in the dataset. The
354 same threshold value is usedfor all nine formulations, so they are comparable in size.
355 The electricity data are not normalized for clustering, because it is the only variable
356 used for the "Electricity" formulation, and it will not change the results.

357 Next, clusters are formed on the social and technical variables independently,
358 and then subsequently the full set of socio-technical characteristics. For these
359 formulations, each of the socio-technical variables are normalized to fall between 0
360 and 1 to avoid any bias in the model towards variables based on the difference in
361 magnitude. For example, "Income" is of the 10^4 order whereas "Female" is of the 10^{-1}
362 magnitude. The algorithm is run 100 iterations per formulation to capture a
363 statistically significant "average" result and to catalogue the variability which exists
364 within the max-p regions solutions for each formulation. The algorithm intentionally
365 does not have a seed, so every run yields slightly different results.

To explore the clustering capabilities on a subset of socio-technical variables, a regression model is run to develop weights based on variable influence to the variation in the electricity data. First regression model with all of the socio-technical parameters is developed and the variables with the highest predictive power are selected to use in clustering. The goal of the regression is to identify significant variables, not to develop a predictive model of electricity use, so regression-specific issues (e.g. multicollinearity, heteroskedasticity) are not corrected. Forward stepwise model building shows similar variable significance levels to running the regression model once with all variables present. Therefore, t-values from the full model are used as a measure of significance (see the Appendix). Once a subset of variables is selected and clustering performed, the variables are narrowed to a smaller subset using the same procedure. This yields three formulations of socio-technical subsets for each case study.

Additionally, the regression coefficients from each subset are used to explore the impact of weighting on the clustering results. The regression coefficients quantify the relative significance of the variables to the variation in electricity use. Including these in the clustering algorithm places more significant variables as higher weighted in the clustering process. For each of the three socio-technical subset models, the max-p algorithm is run another 100 iterations, respectively, while weighting the normalized variables with the regression coefficients. Essentially, this increased the magnitude of significant variables while decreasing (relatively) the magnitude of less-significant variables.

To compare the clustering outputs of each of the nine clustering set-ups in both case studies, several benchmarks and measurements of homogeneity are developed. For the benchmarks, the clustering on only the electricity data provides a lower bound on the variability which exists in the data. As a point of comparison, all clustering statistics on the pre-defined census tracts (CT) are calculated for each city. Comparing variability to CTs indicates whether there is value in defining new geographies for electricity using socio-technical variables, or whether the existing geographical definitions have lower or comparable homogeneity.

3 Results

Using max-p regions clustering with socio-technical characteristics significantly reduces the heterogeneity of electricity consumption within regions compared to using pre-defined CT geography. The reduction is greater for NYC, which has more natural variability in the electricity data compared to LA, where there is a fairly narrow range of variance between households. Weighting the clustering variables based on relative significance does not yield more homogenous geographies in the two case studies; it does however lead to more consistent clustering results between iterations due to the additional constraint of the weights.

3.1 Residential Energy Overview

LA and NYC have distinct climates and building stocks, which impact how each city uses electricity. These differences in climate, infrastructure, and demographics lead to different end use energy consumption profiles for LA and NYC. According to the 2009 RASS, the majority of energy use (~68%) in the LADWP service area is natural gas for space and water heating. The main driver of end-use electricity consumption is

411 the miscellaneous category, which includes interior and exterior lighting as well as
412 items not covered specifically in the energy consumption survey (Figure 3). The
413 accompanying documentation for the RASS analysis states that they believe lighting to
414 be the largest contributor to this category [21]. Next the main contributors are
415 TV/electronics, refrigerators, and heating and cooling. The total energy devoted to
416 heating and cooling in the LADWP area is approximately 36%. Pools (in the form of
417 water circulation and heating) are also a non-negligible contributor to the energy
418 consumption profile
419

Figure 3: Residential end use consumption in LA and NYC. Based on the 2009 RASS and RECS Surveys, residential end use consumption in the LADWP service area and total energy use in NYC is broken down. In LA, most energy comes from electricity so the profiles are similar. In NYC, a large portion of energy comes from non-electric heating, so the profiles show substantial variation.

There are fewer categorical distinctions of end-uses of energy in RECS than are present in RASS. Within RECS, fewer end-use categories are provided than in RASS, so the leading electricity consumer is "Other". This category, presumably, includes consumers such as pools, electronics, small appliances, cooking, laundry, and lighting. Without more data it is difficult to make that distinction. The next largest consumers

of electricity in NYC are refrigerators, followed by heating and cooling. Although many categories are aggregated together in the RECS data, the profile of electricity use is quite similar to LA, since the majority of heating is non-electric (~70%) [22]. This electricity consumption profile is likely to be quite different for areas with heavy (electric) cooling loads.

Total energy consumption shows a slightly different profile than electricity consumption in NYC. Of the non-electric energy sources such as natural gas, fuel oil, and wood, nearly 60% goes to space heating. This significantly influences the total energy profile, making space heating the number one end-use consumer of residential energy. This is again followed by the ambiguous "Other" category, and then water heat, which also uses a lot of non-electric energy. Comprehensive strategies for examining energy use in NYC should therefore consider all sources of energy use because of the heavy heating and cooling loads. There's a lot of potential for reduction in these areas.

3.2 Cluster Heterogeneity

Measures of variability are developed to compare the heterogeneity of different clustering formulations and iteratively run the algorithm to explore the variability within the clustering solution sets. Within each cluster, a weighted (by number of accounts) standard deviation of the average consumption of each CBG within the cluster is calculated. Based upon the standard deviation of each cluster, an average standard deviation for the entire study area is then calculated as a measure of the typical heterogeneity from that clustering run. Likewise, the standard deviation of the heterogeneity measure from each cluster can be measured. Because

multiple clustering solutions exist under each max-p formulation, 100 iterations are then run and the heterogeneity measure and standard deviation of the heterogeneity measure across runs are averaged. These are referred to as the formulation heterogeneity and the formulation standard deviation, respectively, for this article. The results for each formulation are summarized in Table 2. For each formulation, a t-test is performed to calculate the probability that the formulation heterogeneity is less than that of the CTs. In every single case, the p-value is much less than 0.0001, indicating a high level of certainty that this is the case.

Table 2: Average formulation heterogeneity and formulation standard deviation. For each of the eight clustering formulations and the pre-defined CTs the measures of heterogeneity and standard deviation are calculated as defined in Section 3.2. The heterogeneity measure is based upon standard deviation within each cluster and represents average variation of annual electricity use (in kWh). The standard deviation of these variances across the city is calculated as well (also in kWh).

	Los Angeles				New York City			
	Heterogeneity (kWh)	St. Dev. (kWh)	t-statistic	p value	Heterogeneity (kWh)	St. Dev. (kWh)	t-statistic	p value
1. Electricity	437.2	13.1	-93.7	9E-99	694.1	8.1	-367.8	6E-158
2. Social	512.0	14.2	-34.0	9E-57	741.6	10.5	-238.3	1E-138
3. Technical	513.8	14.6	-31.9	4E-54	708.2	8.6	-330.5	9E-153
4. Socio-Technical Full	500.6	11.3	-53.0	9E-75	728.4	8.6	-308.0	1E-149
5. Socio-Technical Subset 1	495.0	15.2	-42.8	6E-66	725.3	7.9	-339.2	7E-154
6. Socio-Technical Subset 2	486.1	14.0	-52.9	1E-74	736.7	8.5	-302.0	7E-149
7. Socio-Technical Full, weights	496.3	12.8	-49.9	3E-72	726.4	9.9	-269.3	6E-144
8. Socio-Technical Sub1, weights	507.3	15.3	-34.7	2E-57	727.2	8.4	-316.1	8E-151
9. Socio-Technical Sub2, weights	505.5	13.1	-41.9	5E-65	733.7	8.6	-299.7	1E-148
Census Tracts (CTs)	560.5	-	-	-	992.8	-	-	-

3.2.1 Los Angeles Results

For LA, a theoretical minimum consumption is established based upon the electricity-only formulation. This formulation has a formulation heterogeneity of 437.2 kWh/household/year, which is approximately 7.9% of average annual household

use in the LADWP service area. This number represents the natural variation which already exists within the CBGs when clustered to this threshold, and represents a pseudo "theoretical minimum" standard deviation which could be achieved through clustering. This is because the most effective way to obtain areas of homogenous electricity use would be to cluster on electricity use itself.

The formulation heterogeneity of the clustering results is compared to pre-defined CT geography to benchmark performance. For the CTs in the LADWP service area there is an average standard deviation of 560.5 kWh/ household/year, which is approximately 10.1 % of average annual household use. This higher variance in the CT geography indicates that clustering reduces heterogeneity. For LA there is not a large difference between the tract variability and the theoretical minimum of the electricity clustering, indicating little room for improvement over the existing geographical definition.

By including the formulation standard deviation, all of the social, technical, and socio-technical cluster formulations outperform the tract geography. In LA, the difference between the different subsets is not significant. The worst performing of the cluster formulations is the technical variables only with just a 0.8% (of average HH use) improvement over the CTs while the best performing is the socio-technical subset A1 without weights, which is the subset that includes just the top six significant socio-technical variables. Weighting the clustering parameters has little impact on the formulation heterogeneity, although it might slightly reduce the standard deviation of the Heterogeneity.

Figure 4: Spatial distribution of CBG deviation from cluster average in LA. This choropleth map shows the difference between the CBG electricity use and the average cluster electricity consumption. Red represents an overestimation and blue represents an underestimation. The maps shows that socio-technical clustering has fewer CBGs with the extreme differences (bright red and blue) compared to the pre-defined CTs.

3.2.2 New York City Results

In NYC both the annual average use and the formulation standard deviations are greater than in LA, with the average residential unit consumption at 6,515 kWh per year, electricity clustering with a formulation standard deviation 694.1 kWh (10.7% of total). Formulation heterogeneity from the CTs is 992.8 kWh, or 15.2% of the average unit use. This is a slightly larger range of potential improvement than exists

for LA, where there is more homogeneity in the residential consumption profile overall.

As in LA, all of the clustering scenarios in NYC were less heterogeneous than the CT geography. The scenario which performed the best consistently was the purely technical clustering with the lowest overall average standard deviation of 708.2 kWh (10.9%). There is little difference between the formulation heterogeneity of the two larger subsets of socio-technical variables (~727 kWh), but the smallest subset actually performed slightly worse with a result of 733.7 kWh. The social clustering showed lower heterogeneity (741.6 kWh) than this smallest subset, but it is not statistically different than the second smallest clustering formulation. In NYC, formulation heterogeneity for total energy use (including fuel) is similarly found to be lowest for technical clustering scenario, and that all clustering formulations improve over the pre-defined CTs (see Appendix A.5). Similar to the LA results, the weighting of the socio-technical clusters did little to reduce the average standard deviation but it did reduce the formulation standard deviation. Perhaps this is best considered a measure of the consistency of the max-p algorithm than a measure of the homogeneity of clusters.

3.3 Socio-Technical Subset Selection

For both stages of the regression selection, parameters that are significant at a 0.1 p-value are kept (see the Appendix). For LA and NYC the significant variables vary substantially, but the presence of a pool and renting the unit are significant to the variation in the consumption of electricity between households in all four clustering sets. As discussed in the introduction, that does not mean that these are the two main

drivers of total electricity consumption. In general, LA tended to have more social variables with statistical significance whereas NYC had a higher amount of technical variables remaining in both subsets (see the Appendix).

3.4 CBGs Variance Distribution

In both cases, clustering creates lower average standard deviation within geographies, and consequentially, a lower difference between each of the CBGs and its cluster average electricity use. In LA when mapping the difference between the clustering formulation with the lowest heterogeneity slight improvements are observed in the deviation of the CBG from its respective cluster (Figure 3). In NYC, mapping the two sets yields obvious variations between the lowest heterogeneity formulation and the CT assignment, implying a substantial improvement in homogeneity of electricity use. (Figure 4).

The difference between the two is more readily seen when the absolute values of the data are plotted against the cumulative percentage of data present (Figure 5). The area between curves represents the difference in the cumulative absolute value of error between the pre-defined CTs and the clustering approach. The greater this area between the curves, the more the algorithm improves over existing geography.

Figure 5: Cumulative deviation of CBGs from cluster average in LA (left) and NYC (right). The deviation of each CBG from the average of its cluster is compared for both the CT assignment as well as the lowest variance Clustering Formulation in each city. Since the goal is to have all of the deviations as close to zero as possible, clustering formulations with a higher cumulative percentage at the lower difference in electricity use perform better.

There are several potential reasons why significant variables in LA include more social variables whereas NYC includes more technical variables. Primarily, LA and NYC have dramatically different heating and cooling loads. It is possible that the base load of electricity required to heat and cool homes is such a dominating amount that what drives the difference in the results therefore is the choice-loads (televisions, pools).

In LA, perhaps there is more choice in the amount of electricity used, so social factors play a larger role.

4 Discussion

The results show that regions defined using the socio-technical building characteristics have lower variance than pre-defined geopolitical boundaries, so a re-consideration of geospatial boundaries used in electricity forecasting is therefore recommended. Furthermore, the use of intelligent spatial aggregation could provide a framework for thinking about the spatial difference in building energy use. In urban-scale building energy analyses, redefining spatial boundaries based upon socio-technical characteristics to compile areas with more similar energy usage than pre-defined geopolitical boundaries provide should be considered. Previous studies show that electricity consumption in the home is correlated to both social and technical characteristics, and this work discovers that using a subset of these characteristics can significantly improve the intraregional homogeneity of energy consumption.

4.1 Patterns of building energy use

There are smarter ways to define analysis boundaries, even without a full understanding of the relative significance of specific socio-technical variables within a local context. The results show that all clustering formulations based on technical, social, or socio-technical variables create areas with more homogenous electricity usage than the pre-defined CTs. The U.S. Census Bureau originally defined geographical sub-boundaries based foremost on existing physical features (e.g. rivers) and political boundaries (e.g. county line), and secondarily based on the socio-economic status of the inhabitants. The intent of CTs is to remain consistent overtime

so that comparative statistics can be calculated, thus with each passing decade and as regions change, census boundaries correspond less to socio-economic status.

The max-p algorithm performs comparably with and without weighting the variables based on their correlation to electricity consumption, suggesting that merely identifying the appropriate variables for clustering is a more important step in defining geographies than understanding their relative importance. To select a solution, the max-p regions algorithm evaluates the difference of the total variance of all variables between regions. This implies that weighting should influence the chosen solution, as a higher weight (for a more significant variable) will increase the variance. Because the weights are based on regression coefficients, the magnitudes are potentially not great enough to significantly influence the variance, or possibly not correcting for positive and negative coefficients keeps this from appearing in the results. Another potential reason that this might not influence the clustering homogeneity of electricity use is the level of multicollinearity in the variables. If one variable is weighted more heavily at the expense of another variable to which it is significantly correlated (e.g. income and floor area), it is possible that the same total variance becomes expressed in the model regardless of weighting since both variables capture essentially the same information. As weighting seems to have no significance in either of the case studies, this suggests that understanding the nuances of variable significance in a specific location is not a hurdle to forming more meaningful geographies for energy analysis.

4.2 Improving Spatial Modeling

Building energy use cannot be isolated from its local context, and analyzing urban energy use at a meso-geography such as the “neighborhood” scale might be more appropriate than building-level or city-wide assessments. As it is defined for this article, the “neighborhood” level is a mesoscale geography that is more resolute than the CTs or clusters in this study and would aggregate households. To manage growing energy demand in the face of restricted or uncertain supply, a system-level understanding of the patterns of energy use that occur across cities is needed, with enough detail to identify the varying amounts of energy being consumed by different types of people at different spatial locations. Furthermore, the practical steps that can be taken to improve existing building stocks need to be investigated, and stocks should be designed with flexibility to meet future, often uncertain, demands. The neighborhood level might be more appropriate than the household level for predicting electricity use as energy consumption depends heavily upon behavior, which can be highly influenced by the activities of others in the same area. Potentially these new neighborhoods, based on both social and technical characteristics, would be the basic unit for a new type of urban energy analysis.

4.2.1 Urban building energy systems

There is potential to improve urban building energy assessment in at least two areas based on the results of this study: 1) re-defining boundaries of meso-assessments and 2) formulating spatially-targeted policies for energy efficiency and conservation. Neighborhood/mesoscale geographies with socio-technical clustering in urban energy analysis can offer many advantages over existing aggregation methods. Foremost, this scale is necessary to provide sufficient detail for understanding

patterns and decision-making. Furthermore, defining mesoscale neighborhoods that can be flexibly changed over time as behavior and use characteristics change will be necessary to monitor how different areas of the city use energy. There also can be advantages in designing studies. If a mesoscale neighborhood has relatively homogenous energy use (based upon clustering), this could aid in limiting data collection to a statistically significant sample from each area. Conversely, if a study starts with full household level data and wants to aggregate (for privacy or data-condensing reasons), areas with similar energy consumption and socio-technical characteristics will be grouped together. Homogeneity beyond just energy use or a subset of the variables might be advantages in developing non-energy metrics as well. Clustering also lowers the intra-cluster variance over pre-defined geographies. Using a pre-defined geography will provide the appropriate *average* statistics for energy use, but it will create areas with higher levels of variance in forecasting, for example, by placing more outliers together.

As building energy technologies continue to become more efficient and increasingly incorporated into building codes, energy conservation and efficiency efforts will be targeted at behavior and other high-energy emerging technologies, and energy clustering can inform a more efficient allocation of these policies. For example, computers and personal electronics have quickly become important contributors to the total residential building consumption profile, and this proportion continues to rise [43]. Electric vehicles are another technology that might dramatically change home energy use profiles. Additionally, new appliances and disruptive technologies may change the way households use energy. Furthermore,

having increased distributed and on-site energy generation will potentially change the way homeowners make decisions about energy use. In this light, conservation and efficiency policies, which often have a large economic cost associated, should be targeted at the areas of buildings and the types of people where they will have the most benefit. Through a redefinition of energy geographies, specific neighborhoods could be identified where, for example, building retrofits will have the largest savings in wintertime heating use. Clustering could provide a way to identify the hotspots of building energy use.

4.2.2 Technological Urban Systems

This study is the first where the max-p regions algorithm has been applied to technical rather than social variables, and there are potential applications for this framework in other urban systems such as water usage, solar panel installation, and information and communication technology penetration. Like energy usage, the function of other urban systems depends on characteristics of the infrastructure and human activities, and using meso-geographies could inform robust studies on these services. For example, if a relationship existed between income, building construction year, and likelihood of solar panel installation at the home, these variables could be used in clustering to identify neighborhoods that are most (or least) likely to install solar panels in the near future. This might have implications for both local grid infrastructure as well as business forecasting.

4.3.3 Survey Design

There are also potential applications of using clustering of electricity use to design smarter surveys of building energy use. This study uses a prototype approach to

approximate the building shell properties and appliance prevalence across two entire cities because of the lack of detailed spatial technical data, but to further understand urban-scale energy use, a more detailed profile of building technical data across cities is needed, at comparable or finer resolution as the social data from the US Census. Performing clustering on a city could be useful for identifying which areas could be considered homogenous and a statistically significant sample could then be collected from each of these regions on technical characteristics. Understanding the building stock as it exists, including how it constrains or permits decisions on energy use will be important for identifying future paths towards lower energy consumption.

5 Conclusions

In this article, it is shown that intra-zone variability of electricity use can be reduced by redefining spatial boundaries with the socio-technical characteristics of homes. This can be done by choosing social and technical variables that have been shown to have high correlations to residential electricity use. A max-p clustering algorithm is used to endogenously determine the clusters in case studies in two US cities: Los Angeles and New York City. Multiple formulations are considered including only technical variables, only social variables, all social and technical variables, subsets of social and technical variables, as well as formulations weighted based on variable significance. To benchmark the performance of the clusters, the intra-cluster similarity of electricity use in homes is compared to a pre-defined geography commonly used in energy analysis: the US Census Tract. For both case studies, all clustering formulations have a statistically significant reduction in consumption variability compared to census tracts, suggesting that redefining boundaries even if

the context-specific drivers (i.e the most important socio-technical variables) are uncertain will improve over existing geopolitical boundaries. This is significant, for multiple reasons. First, this study might provide the foundation of a new framework for urban energy assessment, based upon clustering, at a sub-urban scale. Even if the exact local relationship between socio-technical variables and energy variables is unknown, in these case studies it is shown that clustering can place homes with similar energy consumption profiles into the same geographies much better than using a pre-defined geography, with a high level of statistical significance. Additionally, there is some evidence to suggest that these intermediate geographies might be more appropriate for urban energy assessment than building-level or city-scale. Second, clustered energy assessment can directly inform spatially-targeted policies within a city by identifying “hotspots” of energy use. This level of nuanced understanding of how homes use energy within a city could lead to better economic allocation of local resources. Understanding spatial differences in energy use might also lead to better design of building technology surveys (i.e. thermal shells and appliances) that provide sub-urban detail. Finally, there are potential applications of this algorithm to other urban technical and infrastructure systems beyond urban energy use. Altogether, the framework offers potential to help us rethink how we approach urban environmental, resource, and economic assessments.

Acknowledgements

We wish to thank Bianca Howard at Columbia University for sharing her modeling work on NYC energy use and Stephanie Pincetl at the California Center for Sustainable Communities at UCLA for providing access to LADWP residential electricity use. This

material is based upon work supported by the National Science Foundation (grant nos. GRFP-DGE 1311230, CMMI 1335640, CMMI 1335640 and WSC 1360509). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation

References

- [1] US Energy Information Administration. Energy Flow Mapping by Sector 2013. http://www.eia.gov/totalenergy/data/monthly/pdf/flow/total_energy.pdf.
- [2] US Energy Information Administration. Electricity Flow Mapping. 2013 n.d. <http://www.eia.gov/totalenergy/data/monthly/pdf/flow/electricity.pdf>.
- [3] US Department of Energy. Energy Efficiency Trends in US Residential and Commerical Buildings. 2008.
- [4] Chan ALS. Investigation on the appropriate floor level of residential building for installing balcony, from a view point of energy and environmental performance. A case study in subtropical Hong Kong. *Energy* 2015;85:620-34. doi:10.1016/j.energy.2015.04.001.
- [5] Pan Y, Yin R, Huang Z. Energy modeling of two office buildings with data center for green building design. *Energy Build* 2008;40:1145-52. doi:10.1016/j.enbuild.2007.10.008.
- [6] Tso GKF, Guan J. A multilevel regression approach to understand effects of environment indicators and household features on residential energy consumption. *Energy* 2014;66:722-31. doi:10.1016/j.energy.2014.01.056.
- [7] Larivière I, Lafrance G. Modelling the electricity consumption of cities: effect of urban density. *Energy Econ* 1999;21:53-66. doi:10.1016/S0140-9883(98)00007-3.
- [8] Howard B, Parshall L, Thompson J, Hammer S, Dickinson J, Modi V. Spatial distribution of urban building energy consumption by end use. *Energy Build* 2012;45:141-51. doi:10.1016/j.enbuild.2011.10.061.

- 749 [9] Anderson JE, Wulfhorst G, Lang W. Energy analysis of the built environment—A
750 review and outlook. *Renew Sustain Energy Rev* 2015;44:149-58.
751 doi:10.1016/j.rser.2014.12.027.
- 752 [10] Ratti C, Baker N, Steemers K. Energy consumption and urban texture. *Energy*
753 *Build* 2005;37:762-76. doi:10.1016/j.enbuild.2004.10.010.
- 754 [11] Shimoda Y, Fujii T, Morikawa T, Mizuno M. Residential end-use energy
755 simulation at city scale. *Build Environ* 2004;39:959-67.
756 doi:10.1016/j.buildenv.2004.01.020.
- 757 [12] Jones P, Patterson J, Lannon S. Modelling the built environment at an urban
758 scale-Energy and health impacts in relation to housing. *Landsc Urban Plan*
759 2007;83:39-49. doi:10.1016/j.landurbplan.2007.05.015.
- 760 [13] Stephan A, Stephan L. Reducing the total life cycle energy demand of recent
761 residential buildings in Lebanon. *Energy* 2014;74:618-37.
762 doi:10.1016/j.energy.2014.07.028.
- 763 [14] Caputo P, Costa G, Ferrari S. A supporting method for defining energy strategies
764 in the building sector at urban scale. *Energy Policy* 2013;55:261-70.
765 doi:10.1016/j.enpol.2012.12.006.
- 766 [15] Heiple S, Sailor DJ. Using building energy simulation and geospatial modeling
767 techniques to determine high resolution building sector energy consumption
768 profiles. *Energy Build* 2008;40:1426-36. doi:10.1016/j.enbuild.2008.01.005.
- 769 [18] Rey SJ, Sastré-Gutiérrez ML. Interregional Inequality Dynamics in Mexico. *Spat*
770 *Econ Anal* 2010;5:277-98. doi:10.1080/17421772.2010.493955.
- 771 [19] Pincetl S, Graham R, Murphy S, Sivaraman D. Analysis of High-Resolution Utility
772 Data for Understanding Energy Use in Urban Systems: The Case of Los Angeles,
773 California. *J Ind Ecol* 2015;In Press. doi:10.1111/jiec.12299.
- 774 [20] US Census Bureau. American Community Survey 2012.
775 <http://www.census.gov/acs/www/>.
- 776 [21] California Energy Commission. California Statewide Residential Appliance
777 Saturation Study: Final Report. Sacramento, CA: 2009.
- 778 [22] US Energy Information Administration. Residential Energy Consumption Survey.
779 2005.
- 780 [23] US Department of Energy. Building America Climate Zones. 2009 n.d.
781 <http://energy.gov/eere/buildings/climate-zones> (accessed March 25, 2015).

- 782 [24] Davis LW. Evaluating the Slow Adoption of Energy Efficient Investments: Are
783 Renters Less Likely to Have Energy Efficient Appliances? Cambridge, MA, USA:
784 2010.
- 785 [25] Brounen D, Kok N, Quigley JM. Residential energy use and conservation:
786 Economics and demographics. *Eur Econ Rev* 2012;56:931-45.
787 doi:10.1016/j.euroecorev.2012.02.007.
- 788 [26] Brounen D, Kok N, Quigley JM. Energy literacy, awareness, and conservation
789 behavior of residential households. *Energy Econ* 2013;38:42-50.
790 doi:10.1016/j.eneco.2013.02.008.
- 791 [27] Kavousian A, Rajagopal R, Fischer M. Determinants of residential electricity
792 consumption: Using smart meter data to examine the effect of climate, building
793 characteristics, appliance stock, and occupants' behavior. *Energy* 2013;55:184-
794 94. doi:10.1016/j.energy.2013.03.086.
- 795 [28] Fritzsche DJ. An Analysis of Energy Consumption Patterns by Stage of Family
796 Life Cycle. *J Mark Res* 1981;18:227-32.
- 797 [29] Gillingham K, Harding M, Rapson D. Split Incentives in Residential Energy
798 Consumption Split Incentives in Residential Energy Consumption. *Energy J*
799 2011;33:37-62.
- 800 [30] Longstreth M, Topliff M. Determinants of Energy Savings and Increases After
801 Installing Energy-Conserving Devices. *Energy* 1990;15:523-37.
- 802 [31] Reiss PC, White MW. Household Electricity Demand, Revisited. *Rev Econ Stud*
803 2005;72:853-83. doi:10.1111/0034-6527.00354.
- 804 [32] Anderson KP. Residential Demand for Electricity : Econometric Estimates for
805 California and the United States. *J Bus* 2014;46:526-53.
- 806 [33] Tso GKF, Yau KKW. Predicting electricity energy consumption: A comparison of
807 regression analysis, decision tree and neural networks. *Energy* 2007;32:1761-8.
808 doi:10.1016/j.energy.2006.11.010.
- 809 [34] Choi IY, Cho SH, Kim JT. Energy consumption characteristics of high-rise
810 apartment buildings according to building shape and mixed-use development.
811 *Energy Build* 2012;46:123-31. doi:10.1016/j.enbuild.2011.10.038.
- 812 [35] Dall'O' G, Galante A, Pasetti G. A methodology for evaluating the potential
813 energy savings of retrofitting residential building stocks. *Sustain Cities Soc*
814 2012;4:12-21. doi:10.1016/j.scs.2012.01.004.

- 815 [36] Jakob M. Marginal costs and co-benefits of energy efficiency investments.
816 Energy Policy 2006;34:172-87. doi:10.1016/j.enpol.2004.08.039.
- 817 [37] Florides G., Tassou S., Kalogirou S., Wrobel L. Measures used to lower building
818 energy consumption and their cost effectiveness. Appl Energy 2002;73:299-328.
819 doi:10.1016/S0306-2619(02)00119-8.
- 820 [38] Costa DL, Kahn ME. Electricity Consumption and Durable Housing: Understanding
821 Cohort Effects. Am Econ Rev 2011;101:88-92. doi:10.1257/aer.101.3.88.
- 822 [39] Los Angeles County Assessor Office. Los Angeles County Assessor Database 2009.
- 823 [40] New York City Department of City Planning. MapPLUTO (release 09c). 2009.
- 824 [41] Duque JC, Anselin L, Rey SJ. the Max-P-Regions Problem*. J Reg Sci
825 2012;52:397-419. doi:10.1111/j.1467-9787.2011.00743.x.
- 826 [42] Rey SJ, Anselin L, Folch DC, Arribas-Bel D, Sastre Gutierrez ML, Interlante L.
827 Measuring Spatial Dynamics in Metropolitan Areas. Econ Dev Q 2010;25:54-64.
828 doi:10.1177/0891242410383414.
- 829 [43] Sanquist TF, Orr H, Shui B, Bittner AC. Lifestyle factors in U.S. residential
830 electricity consumption. Energy Policy 2012;42:354-64.
831 doi:10.1016/j.enpol.2011.11.092.

832

833