Title

Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models

Author

Satre-Meloy, Aven

Publication Date

2019-05-01

DOI

10.1016/j.energy.2019.01.157

Peer reviewed

Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models

Aven Satre-Meloy

Environmental Change Institute, University of Oxford, South Parks Road, Oxford OX1 3QY, UK

Abstract

Achieving further reductions in building electricity usage requires a detailed characterization of electricity consumption in homes. Understanding drivers of consumption can inform strategies for promoting conservation and efficiency. While there exist numerous approaches for modeling building energy demand, the use of regularization methods in statistical models can address challenges inherent to building energy modeling while also enabling more accurate predictions and better identification of variables that influence consumption.

This paper applies five regularization techniques to regression models of original survey and electricity consumption data for more than one thousand households in California. It finds that of these, elastic net and two extensions of the lasso—group lasso and adaptive lasso—outperform other approaches in terms of prediction accuracy and model interpretability. These findings contribute to methodological approaches for modeling energy consumption in buildings as well as to our understanding of key drivers of consumption. The paper shows that while structural factors predominate in explaining annual electricity consumption patterns, habitual actions taken to save energy in the home are important for reducing consumption while pro-environmental attitudes and energy literacy are not. Implications for improving building energy modeling and for informing demand reduction strategies, are discussed in the context of the low-carbon transition.

Email address: aven.satremeloy@ouce.ox.ac.uk (Aven Satre-Meloy)

Keywords: Residential electricity demand, Multiple linear regression, Regularization, Dwelling characteristics, Occupant factors, California

1. Introduction

The U.S. is the second largest energy market and emitter of greenhouse gases (GHG) in the world after China, and buildings hold the largest share of U.S. energy consumption at 41%, more than half of which comes from the residential sector [1]. Residential energy demand has remained relatively stable since 1990, yet the sector still accounted for 20% of CO_2 emissions from fossil fuel combustion in 2015. 68% of these emissions were attributable to electricity consumption for lighting, heating, cooling, and operating appliances, with the remainder due to consumption of other fuels for heating and cooking [2]. Electricity demand is projected to grow over the next thirty years, in part because of an increase in the adoption of cooling technologies due to future warming [3, 4].

Demand reduction is believed to play an important role in the effort to reduce emissions from the residential building sector. The U.S. Environmental Protection Agency (EPA) predicts that demand-side efficiency and saving measures could result in a net cumulative demand reduction of 7.83% by 2030 [5]. These measures are also some of the most cost effective for reducing emissions from the power sector, delivering savings at a fraction of the retail cost of electricity [6]. Only recently have demand-side strategies begun to receive closer scrutiny in national and global scenarios for limiting warming to the 1.5°C target agreed in Paris [7]. This emerging literature stresses the importance of demand-side measures for achieving ambitious climate goals and delivering societal co-benefits for health, equity, and security [8].

The residential building sector's large share of electricity consumption and sizable potential for reducing emissions warrant detailed investigations into the drivers of consumption. A deeper understanding of the characteristics of electricity consumption in homes can inform strategies and policies for promoting conservation and efficiency. This is especially important given that much of the existing quantitative research on building energy consumption and prediction has focused on non-residential buildings [9].

Approaches to investigating drivers of building electricity consumption have proliferated in recent years alongside a similar expansion in available data to analyze these drivers. Both statistical and engineering techniques are increasingly applied to diverse, multivariate data to quantify the contribution of different factors to household electricity consumption. These methods benefit from improved computing power, access to large datasets, and new algorithmic approaches for modeling electricity consumption.

Yet despite their proliferation, statistical and engineering methods for modeling building energy consumption face numerous challenges, especially in the context of informing policy development. Hsu [10] summarizes key challenges that are shared across energy analysis research, and several of these are highlighted here.

First the number of factors that possibly influence energy consumption, including structural factors, such as physical dwelling characteristics and efficiency standards, as well as economic, social, and behavioral dimensions, is almost limitless. Understanding the comparative contributions of these different factors to consumption patterns can improve intervention efforts to promote conservation. Second, although the availability of data is improving, it is still difficult and expensive to gather comprehensive data on these factors, so results are often based on small datasets specific to particular geographic, economic, and social contexts. Third, statistical models based on small samples often do not have high out-of-sample predictive accuracy. Especially when the set of possible predictive factors is large (and in 'high-dimensional' problems, larger than the number of observations), models often 'overfit' the data, meaning they do not generalize well to new data and lead to poor predictions and inferences. Fourth, including a large number of predictors in statistical models increases the likelihood of multicollinearity, where multiple predictors have high degrees of pair-wise correlation, which can inflate the standard errors of coefficients in statistical models and lead to misinterpretation [11]. Finally, an additional challenge is the prevalence of missing data, which is common in datasets pulled together from numerous sources, especially from household surveys where completion is not mandatory. Missing data, if not handled properly, can result in loss of information and introduce bias [12].

Overcoming these analytical challenges is important for interpreting model results accurately and properly informing strategies for delivering energy savings, but many of these issues are not well addressed in the energy consumption literature, and statistical techniques to handle these challenges are rarely applied in empirical energy consumption studies [10]. As the following review of literature will show, numerous modeling techniques exist to estimate residential energy consumption, but many of these are geared toward improving predictive performance without also yielding interpretable results. This phenomenon has become increasingly common with advanced machine learning approaches, especially those in the field of deep learning. While improvements in prediction are certainly important for numerous purposes, developing better solutions for reducing energy demand require interpretable models that identify important factors explaining consumption. Thus, statistical approaches that can improve predictive performance while also ensuring a more robust variable selection process are especially relevant for residential electricity consumption research..

This paper therefore makes two primary contributions. First, it contributes to the literature on model selection for electricity consumption by applying regularization techniques to linear regression models of annual electricity consumption. Following Hsu's introduction of these techniques to the energy consumption literature several years ago [10], they continue to be seldom-used despite their demonstrable benefits for improving statistical models and identifying key variables. This paper will show how the use of regularization techniques should be guided by the analysis objective and the structure of the data. It shows how several recent extensions to these methods can improve results for prediction and interpretation objectives when the data contain many different types of variables, which is common in residential energy demand research. The second aim of this paper is empirical, demonstrating the use of these techniques on an original dataset of annual electricity usage data and a wide range of structural and occupant factors for over 1,000 households in California.

The paper is organized as follows: Section 2 reviews related work, both on statistical modeling of energy consumption in buildings as well as on determinants of consumption. It highlights areas of uncertainty and gaps in our knowledge. Section 3 describes the use of regularization methods, including several recent extensions, and the statistical motivations for the modeling approach undertaken in this paper. Section 4 describes data collection, organization, and preprocessing procedures. Section 5 presents results. Implications for both modeling and policy are discussed in Section 6, and Section 7 concludes with a discussion of how the methods used in this paper can inform further research in building energy consumption analysis.

2. Related work

This review of related work is split into two sections. Section 2.1 describes the high-level taxonomy of approaches for building energy consumption modeling and then provides a more detailed review of statistical methods and several key issues that are present, including the competing aims of prediction and interpretation and the need for robust variable selection techniques. Section 2.2 reviews the literature on determinants of household electricity consumption.

2.1. Approaches for modeling building energy consumption

Swan and Ugursal [11] review residential energy consumption models and show that several approaches are appropriate, depending on the scale of interest. These approaches are either top-down or bottom-up, and Figure 1 shows the methods common to each. Top-down models use large, statistical databases to quantify regional or national energy supply requirements. Econometric models use macroeconomic indicators, such as price and income, whereas technological models generally use characteristics of the entire housing stock, such as appliance ownership. These models are useful for predicting trends in consumption for national planning purposes, but they require little detail beyond these broad indicators and thus provide limited insight into the micro-scale factors that influence consumption, including occupant behavior.

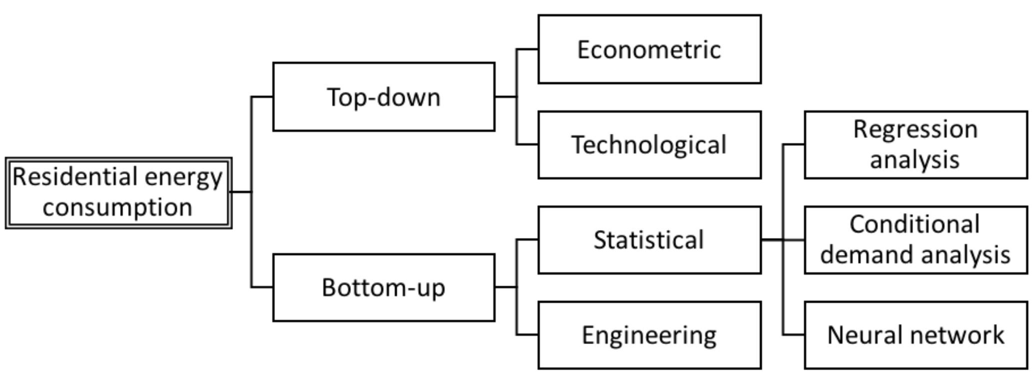

Figure 1: Modelling techniques for estimating residential energy consumption. Adapted from [11].

Bottom-up models, on the other hand, account for energy consumption due to individual end-uses and can use a variety of input data. These data can include socio-demographic, occupant behavior, or technology factors. There are two distinct categories of bottom-up models, which use different approaches for estimating consumption. Engineering methods, also called building physics models, use detailed data on dwelling characteristics, power ratings and use of appliances, and thermodynamic principles to predict consumption. Statistical methods instead use mathematical principles to describe the relationship between predictive variables and household electricity consumption.

The benefits of engineering methods include the use of physically measurable data to determine the consumption of specific end-uses and technologies. Measurements and simulations are useful for describing existing technologies in greater detail and modeling the prospective impact of new technologies. The drawback of using these models is that they rely on assumptions about occupant behavior, do not include other socio-demographic or economic data, and usually require a lot of technical data and measurements of building characteristics while requiring more computational power for analysis [13].

Statistical methods, on the other hand, can incorporate more varied socio-demographic and behavioral data, are often less computationally intensive, and are somewhat easier to develop and use. Several exceptions include nonlinear models, which are discussed in greater detail below. Given that statistical models represent a purely mathematical relationship between energy consumption and predictive variables, however, they are often prone to more error and uncertainty than engineering models [11, 14]. Given recent advances in statistical modeling, and given that statistical modeling techniques are employed in this paper, a brief review of these is provided in the following section.

2.1.1. Statistical and data-driven models

The main approaches for statistical modeling highlighted in Swan and Ugursal [11] are regression analysis, conditional demand analysis (CDA), and artificial neural networks (ANN). More recent reviews include additional methods such as support vector machines (SVM) and decision trees (DT) [15, 9, 16]. Each of these are briefly described in turn. For a more complete review of these methods and their mathematical properties, see Wei et al. [16].

Regression analysis is one of the most common approaches for modeling building energy consumption. In its simplest form, regression analysis determines the size and direction of associations between predictive factors and electricity consumption. Predictors are selected based on expectations of what drives consumption and data that is available or collected. Selecting predictors is the subject of a broad statistical literature, which is further discussed in Section 2.1.2. Models are evaluated using goodness-of-fit measures and model predictive error. Key predictors and their coefficients are examined to determine the strength and statistical significance of their relationships with consumption. Regression models are simple to develop and use, yet they require access to large sets of historical data and do not often achieve the predictive accuracy of other methods.

Conditional demand analysis (CDA) uses regression analysis but only includes as predictors the various end-use appliances owned in the dwelling. The coefficients in the model thus represent the use level and rating of the appliances. While this technique is relatively simple to use, it requires detailed data on household appliance ownership and a large sample of dwellings [11].

Artificial neural networks (ANNs) have grown in popularity with the rise of machine learning disciplines and, especially, deep learning approaches. The method is based on analytic techniques originally developed for studying human neurophysiology. The simplest ANNs include three layers: an input layer, a hidden layer, and an output layer, each of which has interconnected neurons that send signals to the neurons in sequential layers using an activation function [9]. The reason ANNs have gained such popularity is their ability to model incredibly complex, nonlinear relationships. The trade-off for this gain in model complexity is that the coefficients in the model do not have physical significance, so interpreting the influence of different factors in neural networks is challenging.

Another popular method in machine learning is the SVM, which also performs particularly well when the relationship between the inputs and the response is nonlinear. Support Vector Regression (SVR) is the application of SVM principles to regression problems. SVR works by mapping data inputs to a higher dimensional feature space using a kernel function and then constructing a linear model that keeps the error within a predefined threshold. It has shown improved predictive capabilities for building energy consumption [14]. An additional benefit of SVMs is that they require fewer parameters and less training data. Like ANNs, however, SVMs are more complex models that suffer from computational inefficiencies, though optimization of these algorithms is the subject of research [e.g. 15].

Decision trees (DT) work by partitioning data into groups based on predefined predictor variables, where each variable represents a root or branch in the tree, and the data is partitioned into smaller groups along the branches. The modeler chooses which variables to use as nodes and can decide where to trim the DT. In this way, a DT visually represents the data partitioning decisions made at each branch, and for this reason DTs are simple to understand and interpret, which is one of their main advantages. They have also proven effective in building energy prediction [17]. With larger numbers of predictors, DTs can become overly complex, but ensemble methods such as Random Forests can help prevent overfitting [18].

These methods are some of the most common for statistical modeling of energy consumption, but there are many others and many variations on each of these. One of the key differences between these methods, however, is whether they are used primarily for prediction of building energy consumption or explanation of factors that influence consumption. Much of the research interest in machine learning methods such as ANNs, SVMs, and DTs is improving the ability to predict consumption. While this is certainly important in building energy research, the pursuit of more accurate predictions can hamper interpretation efforts. Increasing gains in model accuracy often relies on increased model complexity, which is commonly the case for ANNs and SVMs. This complexity makes it difficult to understand the relationships between data inputs and the response. While regression models may not match the predictive accuracy of complex, nonlinear models, they are interpretable and can clearly describe relationships between variables and energy consumption. When this is the aim, model simplicity is essential. This gives the model practical significance for informing strategies to reduce demand.

For the interested reader, an insightful essay comparing the objectives of prediction and explanation in statistical models is given by Shmueli [19]. The essay concludes that while these objectives often delineate the choice of variables, methods, and approaches for selecting, validating, and evaluating statistical models, in most cases it is appropriate to consider both the predictive and explanatory power of models. Even when the objective is not primarily prediction, the predictive qualities of a model should be reported in research, and *vice versa*. Model performance can then be judged based on both of these criteria.

2.1.2. Variable selection and related challenges in statistical models

Variable selection is a key issue in statistical modeling, especially when the number of candidate variables is large. For data with p variables, the number of possible models with subsets of the p variables is 2^p. A seemingly simple dataset with 10 variables gives over 1,000 possible models with subsets of variables. Even with modern computing capabilities, constructing every possible model and comparing each using some evaluative criteria is impractical as the number of candidate variables grows.

Two additional challenges are more likely to be present when the number of candidate variables increases. The first is multicollinearity between predictors. Multicollinearity exists when one predictor variable has a near linear relationship with another [20]. When this is the case, the coefficient estimates for the regressors become unstable and are susceptible to erratic changes with small changes to the model or data. Multicollinearity has been identified as a challenge in energy consumption modeling in numerous instances [11, 21, 22]. It is a challenge unique to the objective of constructing explanatory models and interpreting variable size and significance, as predictive accuracy does not suffer when multicollinearity effects are present [19].

The second challenge for models with many potential predictors is overfitting. Overfitting occurs when the model is overly complex or includes more variables than necessary. In high-dimensional cases, where models have more predictors (p) than observations (n), approaches such as ordinary least squares (OLS) regression do not have well-defined solutions. The result of overfitting is a model that fits so well to the existing data that it does not generalize to new data. When models are overfit, they suffer from high variance, meaning they capture the noise inherent in the data along with the underlying patterns. High bias models, on the other hand, are too simple and do not fit well to the existing data. There is a well-researched trade-off between bias and variance in the statistics literature [23]. In the energy modeling literature, efforts to address overfitting are most common in predictive modeling or forecasting studies (e.g. [24, 25]).

These challenges stand out in efforts to model energy consumption, especially for explanatory purposes, because of the sheer number of potential factors that influence usage and because of their potential for high pairwise correlation. Certainly, domain knowledge and previous research should guide the selection of relevant variables, but analyses that explore large sets of untested variables are also valuable, and statistical techniques that can address the challenges of large predictor sets, multicollinearity, and overfitting can aid in variable selection efforts. For this reason, there is a rich and active literature in statistics on variable selection [26].

Some of the most popular statistical techniques for selecting variables fall under the stepwise family of approaches, which includes forward selection, backward elimination, and stepwise selection [26]. These procedures iteratively construct regression models by adding or removing predictors based on a test statistic or minimizing an evaluative criterion, such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC), until a final model is attained. Stepwise regression techniques have been applied in numerous studies of energy consumption to identify relevant predictors [27, 28, 29, 30]. Other approaches for variable selection in energy consumption studies include principal components regression (PCR) and partial least squares regression (PLSR) [31, 32, 33].

Stepwise regression as an approach to variable selection has been derided in the statistics literature for violating statistical theory and causing important practical consequences for analysis [34, 35]. Some of these issues include R-squared values and regression coefficients that are biased on the high side, severe problems handling multicollinearity, and predicted values that are falsely narrow. PCR and PLSR do not have these same issues but do present challenges for interpretation because they transform predictor variables into linear combinations of the original predictor variables.

A separate class of variable selection techniques that can address many of these issues is regularization. Regularization methods, also called penalized regression methods, have received substantial attention in statistical research [23], but their application to statistical modeling of energy consumption is still surprisingly rare. When Hsu [10] first showed how the application of regularization methods could improve efforts to identify key factors influencing consumption, his review of three prominent energy journals (*Energy*, *Energy Policy*, and *Applied Energy*) showed only a few papers applying these methods, mostly in economic analyses. An updated search in these journals confirms they continue be seldom used. Fewer than a total of 20 papers in these journals (including *Energy and Buildings*) use regularization methods in modeling energy consumption, and much of their use is concentrated in recent machine learning analyses [36, 37] or in energy forecasting studies [38, 39, 40, 41]. In two cases, these techniques have been used to analyze drivers of residential energy consumption in the U.K. and France [22, 42].

Regularization methods are primarily used to prevent overfitting, but in some cases they are appropriate for handling multicollinearity and also variable selection. They also have consistently shown improved predictive ability in statistical models because they sacrifice some model bias for a sizable reduction in the variance of predicted values. A full description of these methods and several recent extensions is given in Section 3.1.

2.2. Determinants of residential electricity consumption

While the previous section provided a review of related work in the energy modeling literature, this section will provide a review of literature investigating determinants of residential electricity consumption.

Jones et al. [43] provided the first systematic review of international research investigating the determinants of electricity consumption and found that at least 62 factors have been studied, but only 20 of these were shown to unambiguously and consistently show a significant positive effect on electricity use. The authors found that the number of papers confirming a positive effect on consumption is much higher than the number showing a significant negative effect. Factors considered in the literature include: socio-demographic, physical dwelling characteristics and appliance ownership, occupant attitudinal factors and energy literacy, and occupant behavior. Each of these factors will be reviewed in turn in the following sections.

2.2.1. Socio-demographic factors

Of the many possible occupant socio-demographic indicators to investigate, most studies focus on gender, age, and number of occupants, household income, and tenure of the dwelling (whether it is owned or rented). Nearly all studies reviewed show that the number of occupants has a significant, positive effect on household electricity consumption [e.g. 44, 32]. The presence of young adolescents tends to amplify this trend [45, 46]. Wiesmann et al. [47] show that per capita electricity consumption is lower in households with more occupants, and Kavousian et al. [48] find that the rate of usage increase slows with every doubling in occupancy.

The gender of the homeowner is not often statistically significant in regression models for household electricity usage, though Brounen et al. [46] find per capita usage to be lower in dwellings occupied by females even after controlling for wealth.

Age of the occupants shows conflicting associations to usage. Several studies find a negative correlation between age and consumption [46, 48, 49], while others find a positive correlation [50, 22, 51]. Researchers attribute these disparities to the fact that, in some cases, older occupants tend to be more aware of their consumption and use fewer electronic gadgets, but, in others, they spend more time in the home and are thus likely to consume more electricity.

Two other socio-demographic indicators that have often shown significant effects on household electricity consumption are income and tenure. The results on household income are also mixed: numerous studies find a monotonic and positive relationship between household income and electricity consumption [44, 46, 52, 53, 54, 55], but others find that the effect is small when controlling for other variables [44, 48, 47].

Home ownership is associated with higher electricity usage in [45] and [47], but it shows no significant relationship in [48].

2.2.2. Physical dwelling characteristics

The size of a dwelling explains a large percentage of the variance in consumption [46, 56, 22, 48, 45], with detached dwellings using more electricity than apartments or flats [22, 48, 45, 55].

Older houses are shown to consume more electricity, likely due to less efficient building fabrics [57, 45], but some studies do not find this effect statistically significant [46, 48].

Even efficiency measures, such as insulation or double-glazed windows, are shown to have mixed relationships with consumption. Some studies find that they do reduce usage [58, 59, 48]; others find no correlation [53] or even a positive correlation [54]. One explanation given is that insulation measures are often correlated with house size and income.

Ownership of air conditioning (AC) significantly and consistently increases electricity usage [53, 32, 60, 61], more so for central AC than window units. Results are sensitive to the climatic conditions where the study took place [62].

Ownership of more appliances generally correlates to greater electricity consumption [44, 22, 47, 43].

Ownership of devices that are intended to save electricity, including programmable and smart thermostats, smart meters and in-home displays, LED lighting, and others, are not as often included in empirical studies. The role of feedback and its affect on consumption is an area of growing interest

[63, 64, 65, 66, 67]. These studies suggest potentially significant savings.[1]

Electric vehicles (EV) are a new class of electricity use and can lead to significant increases in household electricity consumption [69]. DOE [70] find that ownership of some EV models can double the electricity consumption of a single-family home.

2.2.3. Occupant attitudinal factors and energy literacy

The literature includes occupant attitudes on care for the environment, concern for climate change, and support for energy conservation and renewable energy. Energy literacy, or the extent to which individuals are familiar with and understand key concepts and issues related to energy, and its relationship with electricity usage has not been studied extensively.

Several studies that measure pro-environmental attitudes by asking respondents to rate their level of agreement with environmental statements find that attitudes cannot explain historical electricity consumption patterns but can explain savings in intervention studies or the occupant's self-reported engagement in energy-saving behavior [71, 72, 73]. Vringer et al. [74] find no significant differences in consumption for groups of households with different value patterns, and Bartiaux and Gram-Hanssen [75] conclude that it would generally be difficult to use attitudes toward the environment to explain differences in electricity consumption between countries.

In the few studies where it is included, energy literacy is not found to significantly correlate to either historical consumption or energy conservation behavior [76]. The National Environmental Education & Training Foundation (NEETF) gave a short energy knowledge quiz to a nationally representative sample of 1,503 Americans to determine the public's basic knowledge of energy issues. NEETF's report claims that "higher levels of knowledge of energy production, consumption, and conservation... have a positive effect on the likelihood of engaging in day-to-day activities that directly or indirectly conserve energy or benefit the environment" [77, p. v]. However, the actual reduction in demand was not measured, leaving a gap in our knowledge of the potential of more energy-informed citizens to reduce demand. Only 12% of Americans passed a basic quiz on energy topics, even though 75% rated themselves as having either 'a lot' or 'a fair amount' of knowledge about

[1] See Ehrhardt-Martinez et al. [68] for a meta-review of 36 energy feedback studies from 1995–2010.

energy. Energy literacy has likely not been included in empirical studies as often as other factors because it is inherently difficult and subjective to measure. Most studies that measure energy literacy rates, including the NEETF study, do so with quizzes that ask questions about how and where energy is generated and consumed.

2.2.4. Occupant behavioral factors

Occupant behaviors influence electricity usage [78], and some studies investigating this relationship conclude that reductions of 10–20% in consumption are achievable by modifying behaviors alone [79].

Studies of conservation behavior generally examine either 'habitual' actions or 'purchasing' activities [80]. Gardner and Stern [81] distinguish between these by specifying the former as 'curtailment' behaviors and the latter as 'efficiency' behaviors. They suggest that efficiency-improving actions yield greater savings than curtailing the use of appliances, lights, or inefficient equipment. The other main difference between the two is that curtailment actions must be repeated continuously over time, whereas efficiency measures need only be taken once or a few times and do not require continuing attention and effort. The authors' list of the most effective behaviors inside the home includes turning down the thermostat during the night and curtailing AC use during the day.

A number of studies find that occupant behaviors are important in explaining usage when controlling for structural elements [82, 83, 84, 85]. Huebner et al. [86] warn that similar findings from their study may not be generalizable.

Long-term curtailment behavior is also measured in Kavousian et al.'s [48] research with inconclusive findings on its impact. They find that, contrary to their expectations, the behavior of 'Purchasing Energy-Star Appliances and Air Conditioners' is positively associated with households' daily minimum electricity consumption. They offer, as a possible explanation, the much-studied 'rebound effect' where increases in appliance or device efficiencies result in increased use of them [87]. They also find that those who report a long-term habit of 'Turning Off Lights When Not in Use' consume more electricity on average. This gap between individuals' intentions is investigated by Kennedy et al. [88], who find that 72% of respondents self-reported a gap between their intentions and their actions related to the environment.

3. Methodology

The above literature review highlights a notable gap in the use of regularization methods for energy use models and inconclusive findings on determinants of consumption. The challenge of variable selection looms large in studies of residential electricity usage, and related analytical challenges present difficulties for model interpretation. This section describes the fundamental regularization methods and their application to multiple linear regression models. It first introduces these methods and then describes the motivations for using regularization in this study. It then describes two recent extensions and how they improve on some of the shortcomings of the original regularization methods. Next, it describes the model training, testing, and validation procedures. Lastly, it describes an important step taken during data preprocessing to address the issue of missing data.

3.1. Regularization: overview and motivation

Regularization methods are known as shrinkage methods because they shrink the coefficients of regression predictors, which trades off a small increase in model bias for a greater reduction in variance. The methods do this by applying a penalty term to the least squares estimator, hence the name 'penalized regression'. In the typical regression situation, we have data (x_i, y_i), $i = 1, 2, ..., n$, where x_i and y_i are the regressors and response for the ith observation, respectively, and where x_j denotes the jth predictor, $j = 1, 2, ..., p$. In OLS regression, we aim to estimate predictor coefficients (β_j) by minimizing the residual sum-of-squares with respect to β:

$$RSS(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{1}$$

Penalized regression methods constrain this optimization problem by adding a penalty term in the estimation of model coefficients. Because the penalties depend on the magnitude of these coefficients, the predictors and response are centered and standardized to have mean zero and a standard deviation of 1.

The three fundamental methods in regularization are ridge regression, developed by Hoerl and Kennard [89], lasso regression, introduced by Tibshirani [90], and the elastic net, introduced by Zou and Hastie [91]. The penalty term in each case is slightly different. In ridge regression, the penalized optimization problem is:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left(RSS(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{2}$$

where $\lambda \geq 0$ is the parameter that controls the amount of shrinkage. As λ increases, so does the penalty, which in ridge regression is the sum-of-squares of the coefficients. For this reason, ridge regression is also called l_2-regularization because it constrains coefficients by their l_2 norm. Penalizing by $\sum_{j=1}^{p} \beta_j^2$ has the effect of shrinking model coefficients but never to zero. An interest in yielding sparse, interpretable models is what motivated the introduction of the lasso, which constrains coefficients by their l_1 norm, and is given by:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left(RSS(\beta) + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \right) \tag{3}$$

In lasso regression, this penalty constraint delivers sparsity, meaning some coefficients are set exactly to zero. In this way, beyond improving prediction, lasso performs variable selection and thus provides a level of interpretability in the model.

The motivation for the third regularization method is due to the behavior of lasso given highly correlated predictors. In lasso regression, the penalty tends to set only one of the predictor's coefficients to zero, and this procedure can yield non-unique solutions as well as poorer predictions when important but correlated predictors are removed from the model. A combination of both ridge and lasso penalties is the elastic net penalty:

$$\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1 - \alpha) \mid \beta_j \mid) \tag{4}$$

where α is an additional tuning parameter that can be tuned to constrain the optimization by both the l_1 and l_2 norms. Eq 4 is a generalized formulation of the three regularization penalties. If $\alpha = 1$, this penalty is the ridge penalty, and if $\alpha = 0$, it is the lasso penalty.

These three methods have properties that make them useful in different situations, so it is important that their use is guided by the objective of the analysis. Several points on the motivations for using regularization in this study are thus provided here.

While all three methods are able to reduce model variance and prevent overfitting, the most important difference between the three is whether or not they give a sparse solution. Both lasso and the elastic net penalties can yield sparsity in the model, whereas ridge cannot.

In addition, because ridge regression penalizes coefficients by adding the sum-of-squares of the coefficients, it penalizes the largest βs more than it does the smaller ones. This is sometimes important when inspecting ridge solutions, as it may be difficult to interpret which predictors are influential in the model. As was mentioned previously, the behavior of the regularization methods when multicollinearity is present is somewhat different. The elastic net is said to have a *grouping effect*, which follows the intuition that highly correlated predictors will likely have similar estimated coefficients, and by combining ridge and lasso penalties, it keeps groups of correlated predictors in the model Zou and Hastie [91]. Ridge regression shrinks highly correlated predictors' coefficients toward one another, but this effect is still preferable to the situation with lasso, which tends to arbitrarily set one of the highly correlated predictors to zero. Elastic net is thus the preferred method when both multicollinearity is present and sparsity is an objective of the analysis. When multicollinearity is not an issue and pairwise correlations between predictors are low, there is often little difference in predictive accuracy between lasso and elastic net.

In high-dimensional situations where $p \gg n$, lasso can at most select n predictors, which was shown by Zou and Hastie [91] to be a limiting feature in variable selection. In these situations, both elastic net and ridge can select more than n predictors and are preferable for more accurate models. Again, the elastic net is preferred over ridge in situations where sparsity is a goal.

The motivations for using regularization in this paper are thus guided by the following characteristics of the data. First, the number of predictors ($p = 60$) is not larger than the number of observations ($n = 1008$), so the data are not 'high-dimensional' even though the predictor set is quite large.

Second, multicollinearity among predictors does not appear to be an issue. Multicollinearity can be investigated by inspecting variance-inflation factors (VIFs), which signal whether regression coefficients are inflated due to correlation between predictor variables; if they are uncorrelated, VIF = 1. Traditionally, VIFs greater than 10 indicate high multicollinearity [92], but recent work suggests that the cut-off point for VIFs should be much lower—Diamantopoulos [93] set the limit at 3.3. The VIFs for the predictors in the present data range from 1.08—2.44 with a mean of 1.48, well below the cut-off point that signals potential issues.

This study has the stated aims of addressing overfitting to improve predictive performance while also performing variable selection to construct a more parsimonious model for interpretation purposes. The data are not high-dimensional (though still include more predictors than would be tractable using best-subsets methods), and pair-wise correlations between predictors are low. For this reason, lasso and elastic net are expected to perform similarly. As the next section will show, however, two extensions of the lasso may be expected to improve on both aims stated in this paper, given several additional aspects of the data.

3.2. Extensions of the lasso: group and adaptive lasso

The previous section discussed some of the situations where lasso regression does not perform adequately, such as when multicollinearity effects are present or in $p \gg n$ situations. An additional challenge for lasso regression is handling categorical predictors. Yuan and Lin [94] showed that lasso is designed to select individual predictors rather than groups of predictors. As categorical predictors are normally coded as multiple dummy variables, where each dummy represents a different category, it makes sense in analysis to consider these variables together rather than separately when applying the shrinkage penalty. It would be inappropriate to include some of these variables in the model but not others. Especially in the building energy domain, categorical predictors are quite common (e.g. type of building, type of heating system, ownership structure).

To address this drawback of the lasso, Yuan and Lin [94] introduced the group lasso (*gLasso*), a generalization of the standard lasso optimization problem. With p predictors divided into L groups, where p_l is the number in group l, and where, for ease of notation, X_l represents the predictors corresponding to the lth group, with corresponding coefficients β_l, the group lasso solves the convex optimization problem:

$$\hat{\beta}^{gLasso} = \underset{\beta}{\operatorname{argmin}} \left(\|y - \sum_{l=1}^{L} \beta_l X_l)\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \|\beta_l\|_2 \right) \tag{5}$$

where $\sqrt{p_l}$ accounts for the varying group sizes, and $\|\beta_l\|$ denotes the l_2 (Euclidean) norm of the coefficients, which is not squared. Thus, instead of constraining the optimization by the sum of the absolute value of individual coefficients, group lasso constrains by the l_2 norm of groups of coefficients. Like in lasso, depending on the value of λ, entire groups of predictor coefficients may be set to zero.

A second challenge with lasso is that variable selection can be inconsistent and that many noise variables can be included in the estimate, especially with increasingly large p. Meinshausen and Bühlmann [95] and Zhao and Yu [96] show that this shortcoming leads to conflict between optimal prediction and consistent variable selection. They show that the optimal λ for prediction can give inconsistent variable selection results, including noise variables in the model and biased estimates for large coefficients. These studies confirm that under certain conditions, lasso does not possess the 'oracle' property. In the context of linear regression, a method possesses the oracle property if it consistently and correctly selects the nonzero coefficients, and their estimates are the same as they would be if the zero coefficients were known in advance [97].

Zou [98] confirm that there are scenarios in which lasso selection cannot be consistent and thus does not possess the oracle property. They propose a new version of the lasso, the adaptive lasso ($adLasso$), which addresses this problem. It does so by including adaptive weights to penalize different coefficients in the l_1 penalty. Estimates for adaptive lasso are given by:

$$\hat{\beta}^{adLasso} = \underset{\beta}{\operatorname{argmin}} \left(RSS(\beta) + \lambda \sum_{j=1}^{p} \hat{w}_j \mid \beta_j \mid \right) \tag{6}$$

where $\hat{w}_j$ is a vector of adaptive weights assigned to the different coefficients. The weights vector is defined as:

$$\hat{w}_j = \frac{1}{(\mid \hat{\beta}_j \mid)^{\gamma}} \tag{7}$$

where $\hat{\beta}_j$ here is an initial estimate of coefficients, usually either $\hat{\beta}^{OLS}$, $\hat{\beta}^{ridge}$, or $\hat{\beta}^{lasso}$, and γ is a positive constant for adjustment of the adaptive weights vector. Zou [98] suggest values of 0.5, 1, and 2. Given this additional weights multiplier, adaptive lasso penalizes coefficients with lower initial estimates more than it does larger coefficients. The authors explain that in $p \gg n$ situations, l_2 regularization can be used to compute the initial estimates of coefficients, given that both l_1 regularization and OLS are not appropriate estimators in high-dimensional settings. This paper uses $\hat{\beta}^{OLS}$ estimates and $\gamma = 1$ for the adaptive weights vector.

3.3. Model training, selection, and validation

All five of these regularization methods are applied to multivariate survey and annual electricity consumption data for a large sample of U.S. households in California. In addition, a stepwise regression method is also applied for comparison purposes. This section explains the procedures to train the models, select models using cross-validation, and test these on hold-out data.

Models for each regularization method are trained on a sample of 80% of the observations, subsequently referred to as the 'training set', with 20% held out as the 'test set'. Motivations for this split and for holding out the test set are given in [23].

Whereas best-subset and stepwise methods use test statistics for model selection, these are less appropriate for regularization methods. Instead, cross-validation is used for model selection. In cross-validation, we fit the model using a sample of 90% of the observations and then use it to predict the remaining 10% of the data in order to obtain the mean-squared error (MSE). This is repeated for k (usually 10) 'folds', and the MSE is averaged over these folds.

In regularization, it is typical to use cross-validation as a means of computing model MSE for a range of different λ values in order to see how increases in the strength of the penalty term relate to trade-offs between the bias and variance of the model. Plotting the relationship between λ and MSE error obtained through 10-fold cross-validation enables the modeler to select a final model that minimizes MSE error or (given the objective of analysis), select a more parsimonious model that still gives an MSE within one standard error of the minimum. This last piece of guidance is given in Friedman et al. [99, p. 17]. In elastic net regularization, both λ and α are tuned simultaneously across a range of values to find the combination of l_1 and l_2 penalties that minimizes the MSE.

After selecting a model for a given value of λ, the model is applied to the test set, and fitted values for the response are compared with actual values to determine prediction error. To evaluate each of the regularization models, we compare three criteria: model root mean-squared error (RMSE), R-squared for the test set, and the number of nonzero coefficients. These criteria permit an evaluation of the competing aims of prediction and sparsity for the models.

There are several reasons why typical inferential constructs such as confidence intervals and p-values are not calculated in this analysis. One reason is that inference is not entirely appropriate given the non-random sample of households studied. Additionally, these inferential constructs do not generally exist for penalized regression estimates. Taylor and Tibshirani [100] state this problem in simple terms: if we use a regularization method for variable selection, we have already searched for the strongest associations in the data and selected these. This means the bar for declaring associations significant must be set higher. There is an emerging literature on post-selection inference [101, 100, 102], but in this paper, given the nature of the sample and the aim to identify and describe factors that have strong associations with electricity usage, inference is not part of the analysis.

3.4. Multiple imputation for missing data

The penalized regression approaches introduced in the previous section can improve the performance of statistical models when many of the challenges discussed are present. The final challenge mentioned in the introduction was that of missing data, which is not addressed through regularization.

Issues of missing data are well-documented and quite common in social science research. Several authors conducted a review of literature employing surveys in political science journals and found that "approximately 94% use listwise deletion to eliminate entire observations (losing about one-third of their data, on average) when any one variable remains missing. . . " [103, p. 45]. Statisticians and methodologists agree that this is a poor approach to handling missing data because it can both result in the loss of information and introduce bias into regression models [12]. In the case of this study, 126 full observations would have been deleted following this approach (a loss of 13% of the data). For this reason, a multiple imputation (MI) method is used to handle missing data.

Multiple imputation (MI) extracts information from the observed variables with a statistical model (for instance, a linear model), uses the model to predict multiple values for each missing data point, and then uses these to construct multiple completed datasets [104, 105]. In each imputed dataset, the observed values are the same while the imputed values vary based on the uncertainty in predicting each missing value. The analysis can then proceed as it normally would on each of these full datasets, afterwards combining or 'pooling' the results.

Improved computational power has made MI relatively easy to implement. This paper uses the Expectation-Maximization with Bootstrapping (EMB) method to create and implement an imputation model with m datasets. For the sake of brevity, algorithmic details are not included here, but they are available in detail in Honaker et al. [106] and Takahashi [107].

Missing values are assumed to be missing at random (MAR), meaning "the probability of missing data on a particular variable may depend on other observed variables (but not itself)" [12, p. 22]. This differs from missing completely at random (MCAR), where missing data are missing due to random error, and not missing at random (NMAR), where missing data are due to respondents refusing to answer questions for specific reasons, and these answers cannot be predicted from the other data. A relevant example is household income, where refusal to answer this question may be non-random. This analysis assumes income can reliably be predicted from other variables in the data, such as age, size of dwelling, and others.

3.5. Software

The statistical software *R Statistics* is used for all analyses [108]. Ridge, lasso, and adaptive lasso are all computed using the *glmnet* package [100]. Group lasso is computed using *gglasso* [109], and elastic net is computed using *caret* [110]. This is also the package used to evaluate final models. *MASS* is used to compute a stepwise regression model for comparison purposes [111]. Finally, *Amelia II* handles multiple imputation [106].

4. Data

The data for this study come from a detailed survey and a database of electricity usage for utility customers in Palo Alto, California. This section introduces the data collection procedures and then presents tables of descriptive statistics for all variables.

4.1. Palo Alto residential profile

Palo Alto is a city in the California Bay Area. It has 66,500 residents, a mild, Mediterranean climate, and an average of 2,832/304 heating/cooling degree days [112]. The city has a target of reducing emissions 80% by 2030 and has already achieved reductions of 36% from 1990 levels [113]. It aims to achieve 16% of these reductions from reducing energy use in existing homes.

Palo Alto's average residential electricity use is 529 kWh per month, similar to the state-wide average of 557 kWh [114] but well below the U.S. average of 900 kWh [115].

4.2. Data collection

A detailed household survey was delivered by e-mail to customers of the city's municipal utility, City of Palo Alto Utilities, which is the sole provider of electric, gas, and water utilities for most of the city's residents. The survey covers 56 questions on occupant socio-demographics, physical dwelling characteristics, occupant attitudes toward the environment, knowledge of energy issues, occupant curtailment behaviors, and energy efficiency program participation. Survey questions were refined with the help of a focus group of the utility's customers.

Utility customers for whom an e-mail address was on record received an invitation to participate. Of 11,963 emails, 4,639 were opened and 1,247 surveys were completed. The completion rate was 8% without incentive, 15% when offered entry into a lighting retrofit lottery, and 27% when offered an LED lightbulb.

Historical billing data for all of the utility's customers was shared with the researcher. Households were excluded from analysis if their home address was incomplete or did not match the utility records (79 cases). Households with PV installations were removed to avoid erroneous use of net-demand data (160 cases), leaving $N = 1,008$ for use in this analysis.

4.3. Independent variables

Independent variables are grouped into categories matching those reviewed in the literature. In the tables below, variables are presented along with their coded numerical ranges and descriptive statistics (where variables are continuous, means are presented as 'M' and standard deviations as 'SD'; for categorical variables, the categories in bold indicate reference categories for regression analyses).

Tables 1 and 2 show the socio-demographic variables and characteristics of dwellings in the sample. Where data is available, these tables also include variable frequencies for the full city-wide population from the American Community Survey (ACS) [116]. Overall, the sample is a good representation of the Palo Alto population, while property owners, elderly households, and detached dwellings are overrepresented.

Energy literacy is assessed with the questions in Table 3. Correct responses to each question are bolded in the table. These items are based on similar work by DeWaters and Powers [117], DeWaters et al. [118], Coyle [119], Brounen et al. [76], and Southwell et al. [120]. On average, participants scored 4 out of 7 possible correct answers.

Description (codes)	Response	Sample frequency	Population frequency
Gender (0–1)	**Female**	39%	49%
	Male	61%	51%
Age range (1–8)	18–25	<1%	29%
	26–35	3%	12%
	36–45	10%	14%
	46–55	22%	15%
	56–65	25%	11%
	66–75	25%	9%
	76–85	12%	5%
	86 and above	2%	3%
Highest level of education obtained (1–3)	Some college or less	4%	20%
	College graduate (four-year degree)	26%	28%
	Postgraduate	69%	52%
Tenure: own or rent (1–2)	**Own**	87%	55%
	Rent	13%	45%
Number of occupants (1–5)		$M = 2.53$, $SD = 1.13$	$M = 2.53$
Total household income before taxes during past 12 months (1–5)	<\$50,000	8%	20%
	\$50,000–\$99,999	16%	18%
	\$100,000–\$199,999	34%	28%
	\$200,000–\$499,999	34%	34%‡
	>\$500,000	8%	
Electric rate schedule (1–2)	**Regular electric**	98%	
	Time-of-use	2%	

Note: Population $N = 66,478$. Population data are from the American Community Survey (ACS) [116].

‡ Includes '\$200,000 and above'.

Table 1: Summary and descriptive statistics for socio-demographic variables.

Description (codes)	**Response**	**Sample frequency**	**Population frequency**
Size range of home in square feet (1–6)	Less than 1000	11%	
	1001-1500	23%	
	1501-2000	29%	
	2001-2500	19%	
	2501-3000	10%	
	More than 3000	8%	
Year of construction (1–3)	Pre-1950	28%	23%
	1950-1989	57%	59%
	1990–present	15%	18%
Type of home (1–2)	**Attached or apartment building**	20%	56%
	Detached home	80%	44%
Number of bedrooms (1–5)	1 or 2	19%	42%
	3	36%	31%
	4	33%	20%
	5 or more	11%	7%
Home has double- or triple-glazed windows (0–1)†	**No**	30%	
	Yes	70%	
Home has floor insulation (0–2)	Not sure	21%	
	No	54%	
	Yes	25%	
Home has roof insulation (0–2)	Not sure	10%	
	No	14%	
	Yes	76%	
Home has wall insulation (0–2)	Not sure	18%	
	No	24%	
	Yes	58%	
Presence or not of an air conditioning system (0–1)	**Does not have AC**	61%	
	Has AC	39%	
Energy devices present in the home (0–1)	Solar water heating	4%	
	LED lighting	77%	
	Smart meter	5%	
	Wi-Fi thermostat	14%	
	Programmable thermostat	58%	
	Plug-in electric vehicle	13%	
	In-home energy display	2%	
	Other energy device	2%	

Note: Population $N = 27,555$ households. Palo Alto data are from the ACS [116].

† 'Not Sure' combined with 'No' responses given low frequencies in these categories.

Table 2: Summary and descriptive statistics for physical dwelling variables.

Energy literacy quiz question	Response option	Frequency
1. Who owns your utility company?	A private entity	0.7%
	The State of California	0.3%
	City of Palo Alto	97%
	Pacific Gas & Electric (PG&E)	2%
2. How much do you pay per kWh for electricity?	Less than 5 cents	6%
	5 - 10 cents	25%
	11 - 20 cents	56%
	21 - 30 cents	8%
	More than 30 cents	5%
3. How much electricity do you think an average Palo Alto single-family household consumes each month?	0 to 10 kilowatt-hours (kWh)	1%
	11-100 kWh	9%
	101-500 kWh	42%
	501-1,000 kWh	41%
	1,001-5,000 kWh	7%
4. Which of the following resources generates the most electricity in California?	Oil	8%
	Coal	5%
	Natural gas	49%
	Nuclear	4%
	Hydroelectric	29%
	Solar	3%
	Wind	2%
5. What percentage of the electricity supplied by City of Palo Alto Utilities is carbon neutral?	20	16%
	30	23%
	50	19%
	70	16%
	100	25%
6. Which of the following uses the most energy in the average Palo Alto home over the course of a year?	Lighting	4%
	Powering household appliances	14%
	Heating water	11%
	Heating and cooling rooms	63%
	Refrigerating food	9%
7. Of the following household appliances, which do you think consumes the most electricity while being used?	Dishwasher	7%
	Fridge/freezer	19%
	Laptop computer	2%
	LED light bulb	1%
	Electric space heater	71%

Table 3: Energy literacy quiz questions and response frequencies.

Table 4 shows the occupant attitude variables and frequencies. These are either measured on a 5-point Likert scale (‘Strongly disagree’ to ‘Strongly agree’) or are dummy coded. The final variable in this section is binary coded and thus measures whether respondents believe renewable energy is beneficial primarily for environmental impact (1) or other reasons (0). The mean correlation coefficient between all attitude variables is $r = 0.17$. The Likert scale questions show slightly stronger correlations, with a mean correlation coefficient of $r = 0.29$.

Description (codes)	**Mean (SD) or Response (frequency)**
Saving energy is important	4.62 (0.64)
I would do more to save if I knew how	3.81 (0.88)
We don’t have to worry about conserving energy because new technologies will be developed to solve problems†	4.17 (0.91)
California should produce more electricity from renewables	4.39 (0.80)
Laws protecting the natural environment should be made less strict to produce more energy†	4.03 (1.08)
The way I personally use energy does not really make a difference to the energy problems in California†	3.74 (1.02)
My decisions to participate in energy efficiency programmes are mostly driven by the amount of money I can save†	2.91 (1.10)
Renewable energy is still too expensive to be practical for California†	3.55 (1.09)
When you think about energy, what are the most important values to you? (0–1)	Comfort (46%) Ease of use (31%) Expense (71%) Safety and security (49%) Ability to go off-grid (7%) Environmental stewardship and protection (67%)
What do you see as the most important benefit of renewable energy? (0–1)	**Reducing impact on environment (80%)** Reducing personal energy costs (7%) Decreasing dependence on foreign energy imports (6%) Helping support ‘green’ job creation (2%) Enabling off-grid capabilities (2%) I do not see any benefits to renewable energy (1%) Other (2%)

† Likert scale is reverse coded.

Table 4: Summary and descriptive statistics for occupant attitude variables.

Behavioral variables are shown in Table 5. Except for the efficiency and rebate variables, which are measured as continuous predictors, these variables are measured on a 3-point Likert scale (‘Never’, ‘Sometimes’, ‘Always’). Correlations are generally low, with a mean correlation coefficient of $r = 0.11$. While the means for the curtailment variables indicate high frequencies of energy saving behavior, especially curtailing AC use, both energy efficiency program participation and rebate uptake are low.

Description (codes)	Mean (SD)
How often do you...	
Turn off lights and electrical appliances when not in use	1.70 (0.47)
Unplug electrical appliances when not in use for an extended period	0.87 (0.71)
Take a shorter shower to conserve energy used for heating water	1.33 (0.66)
Purchase appliances that are ENERGY STAR® or energy efficiency labeled	1.58 (0.56)
Only run the dishwasher or clothes washer/dryer when full	1.75 (0.50)
Turn down thermostat while asleep in the winter†	1.76 (0.54)
Turn off AC when no one is home in the summer†	1.86 (0.37)
Talk with other members of your household about your energy bill†	1.49 (0.72)
Talk with your friends or neighbours about your energy bill	0.78 (0.76)
Talk with your friends or neighbours about ways to conserve energy	1.02 (0.76)
Talk with your friends or neighbours about your own energy efficient devices or technologies	1.01 (0.78)
Number of energy savings programmes respondent participated in (0–3)‡	0.63 (0.82)
Number of energy rebates respondent has received (0–3)‡	0.40 (0.74)

† N/A response frequencies (*TurnDownTherm* = 29; *TurnOffAC* = 618; *TalkAboutBillFam* = 79).

‡ Includes '3 and above'.

Table 5: Summary and descriptive statistics for occupant behavior variables.

4.4. Missing data

Table 6 shows the frequency of missing data for the socio-demographic variables, which were made optional on the survey. Income has the most missingness, while missingness amongst the other data is generally low.

Variable	Missing (%)
Gender	2
Age	2
Education	1
Tenure	1
Occupancy	1
Income	13

Table 6: Missing value frequencies for socio-demographic variables ($N = 1,008$).

After specifying these variables and setting their logical bounds from the variable codes, MI using the EMB method described in Section 3.4 was used to impute five completed datasets.[2] Plots for each of the socio-demographic variables across these five sets were inspected and compared with the original data, which show similar distributions, thus providing a degree of validation. Distributions for the income variable across the five imputed datasets can be found in Appendix A.

Again, because listwise deletion would reduce the number of observations by 13%, the goal for imputation is to avoid losing this important information when conducting subsequent analyses. The total missingness of the data is low, however, so the additional step of 'pooling' results across the five imputed sets is not taken due to computational complexities. Instead, one of the five imputed datasets is randomly selected and used in all subsequent analyses.

4.5. Dependent variable: Annual electricity consumption

The dependent variable for the regression analyses is 2016 annualized electricity consumption in kilowatt-hours (kWh). Table 7 shows electricity usage summary statistics for both the sample and the utility's full customer population. The sample includes 12 households with more than 18,000 kWh

[2] The authors of *Amelia II* recommend a standard value of $m = 5$ [106].

for the year. The correct operation of their meters was validated, and they are kept in the sample.

	N	Mean	SD	1st Quantile	Median	3rd Quantile
Sample	1,008	6,116	3,656	3,759	5,449	7,585
Population	20,006	6,040	5,596	3,130	4,930	7,430

Table 7: Electricity usage summary statistics for sample and customer population.

To address the heteroscedasticity of regression errors, the dependent variable is log-transformed prior to analysis. While the log-transformed electricity usage distribution still exhibits some skew, it is more normally distributed. The log transformation is chosen to improve the regression residuals while still enabling a relatively simple interpretation of results.[3] The sample's mean electricity consumption before transformation is $M = 6,166$ with a standard deviation of $SD = 3,656$. Considering the wider California Bay Area, the mean annual electricity consumption across eight Bay Area counties in 2015 was 6,096 kWh [114, 116].

5. Results

The five regularization methods introduced in Sections 3.1—3.2 are applied to the dataset of household survey responses and log-transformed annual electricity consumption. A stepwise regression is also computed to provide some comparison between regularization and other variable selection techniques. The total number of predictors included in the data is 58.

Figures 2–3 show the results for 10-fold cross-validation to tune the penalty parameter λ and select an optimal model using each regularization method. These plots show how cross-validation MSE varies as a function of the penalty parameter. High bias models are expected on the right side of these plots where the values of λ are higher, whereas high variance models are expected

[3]The dependent variable changes by 100 $\times$ (coefficient) percent on average for each one unit increase in the predictor variable while all other predictor variables are held constant. If the predictor is a dummy variable, when its value switches from 0 to 1, the percent change of the dependent variable is $[100(e^{B_1} - 1]$ while the reverse is $[100(e^{-B_1} - 1]$, where B_1 is the predictor's coefficient [121].

on the left side where the values of λ are lower. In the cases shown here, the characteristic U-shape of the the bias-variance trade-off is very slight (and in some cases absent altogether). This suggests the models are not overfitting much, even with very small amounts of regularization. This may be due to the number of predictors being large but not in comparison to the number of observations. The plots do, however, show that models with heavy penalties have high bias and greater cross-validation MSE as a result.

The plots also show that there is not a sizable difference in the regularization paths for the five methods, and each is able to achieve similar minimum cross-validation MSE (albeit at different strengths of the penalty parameter).

The main difference between the methods, then, can be seen in their sparsity or number of nonzero coefficients, which is indicated along the top horizontal axis. While ridge regression does not set any variable coefficients to zero, retaining all 58 predictors in the final model, both lasso and elastic net achieve similar levels of sparsity, though elastic net reaches a model with minimum MSE needing 20% fewer predictors than lasso (left vertical dotted lines). For the most sparse models that have a cross-validation MSE within one standard error of the minimum (right vertical dotted lines), elastic net and lasso methods both select 21 predictors, which is a 60% reduction from the original total.

Group and adaptive lasso select even more parsimonious models. The plots in Figure 3 show that group lasso finds a model within one standard error of the minimum containing 16 predictors, while adaptive lasso selects a model containing just 11 predictors, a reduction of 81% of the original predictor set.

In order to compare the performance of these methods with other variable selection approaches, a forward stepwise regression is computed using AIC as the criteria for model selection. Next, each of these six models is applied to the test set. For all regularization models, the model within one standard error of the minimum is the model used on the test data. The rationale for this is that selecting the most parsimonious model across each method permits comparisons between predictive error and model interpretability, which is the key objective of this analysis.

Models are compared across several criteria, including root mean-squared error (RMSE) and R-squared for predictions given the test data, as well as the number of nonzero coefficients in the model. RMSE is measured in units of the dependent variable, in this case log-transformed annual electricity consumption, which has a mean of 8.57 and a standard deviation of 0.56. Table

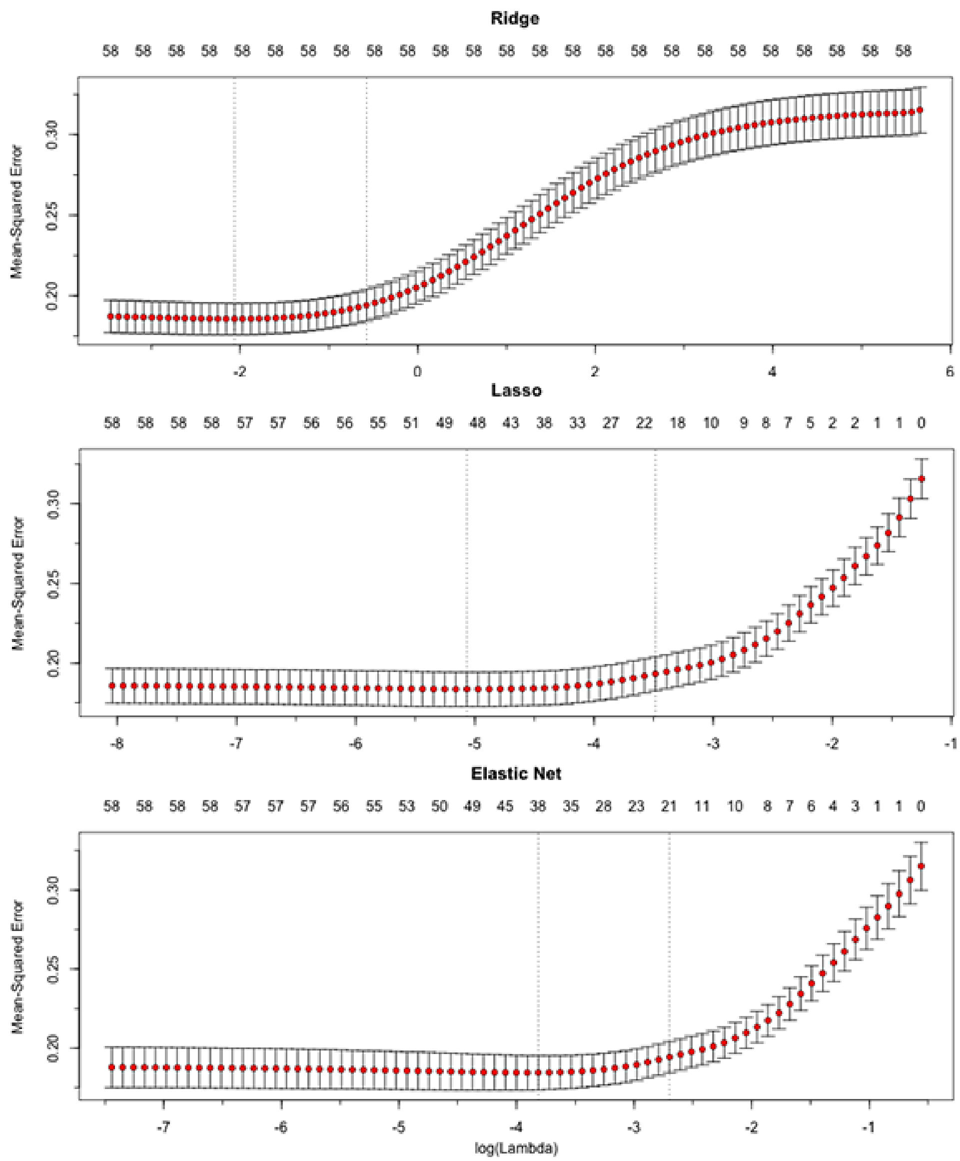

Figure 2: Plots of cross-validation MSE for ridge, lasso, and elastic net models. The horizontal bottom axis shows the logarithm of the tuning parameter λ, while the top horizontal axis shows the number of nonzero coefficients in each model. Points and error bars represent the mean and standard error of cross-validation MSE, respectively. The vertical dotted lines give the model with the minimum MSE (left) and with the fewest nonzero coefficients within one standard deviation of the minimum MSE (right).

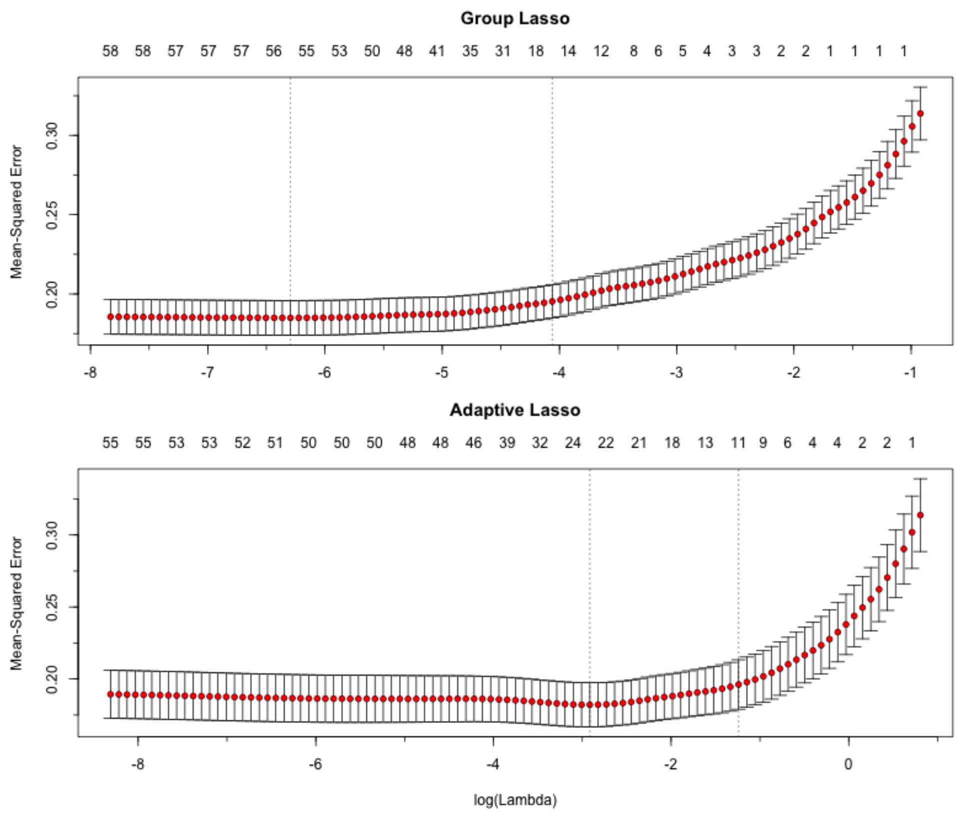

Figure 3: Plots of cross-validation MSE for group and adaptive lasso. Axes and plot elements are the same as in Figure 2.

8 shows results for all six methods, ordered by increasing RMSE. The table shows that elastic net and group lasso achieve the lowest test data RMSE. The stepwise method gives a lower test set error than either adaptive lasso or lasso, while ridge gives the highest error. In general, however, the errors are narrowly distributed, signaling not much difference between methods for prediction (which is similar to the results from cross-validation). Note that R-squared is not adjusted, meaning this value does not take into account the number of predictors in the model. Those models with more coefficients would have lower adjusted R-squared values.

Variables selected by elastic net, group lasso, and adaptive lasso and their standardized coefficients are shown in Table 9. Both elastic net and lasso select the same predictors without much variation in their coefficients,

Method	RMSE	R-squared	Nonzero Coefficients
Elastic Net	0.4457	0.3959	21
Group Lasso	0.4475	0.3893	16
Stepwise	0.4481	0.3978	27
Adaptive Lasso	0.4535	0.3789	11
Lasso	0.4538	0.3733	21
Ridge	0.4562	0.3751	58

Table 8: Model root mean-squared error (RMSE), R-squared, and number of nonzero coefficients for methods applied to test data.

while both group lasso and adaptive lasso achieve more parsimonious models, which is why variables in these models are inspected. The table separates predictors of high and low usage and lists these in order of standardized coefficient magnitude (averaged across the three model selection methods). Unstandardized coefficients are measured in the original units of each independent variable and thus cannot be accurately compared with coefficients of other independent variables measured on different scales. Because most of the variables included in this model are measured in different units (e.g. *Age* in years and *Income* in dollars), standardized coefficients allow for a comparison of each predictor's relative importance in explaining household electricity consumption.[4]

The positive predictors selected by all three methods are EV ownership, size of home, home occupancy levels, type of home, and AC ownership. Other positive predictors include valuing comfort in relation to energy use, household income, solar water heating, being on a time-of-use rate, and respondent age.

The frequency with which respondents report turning off AC when not home, unplugging appliances when not in use, and taking a shorter shower to save energy are associated with lower use, as is renting versus owning a home. Because less than half of the study sample reported ownership of AC,

[4]Standardized regression coefficients are measured in standard deviations rather than in the original units of the independent variable, so the coefficient indicates the number of standard deviation changes expected in the dependent variable for a one standard deviation change in the independent variable.

to examine the effect of the AC curtailment variable, the three methods are used to fit models to survey data for only those households that own AC ($N = 390$). Both elastic net and adaptive lasso select models including the variable measuring AC curtailment. Cross-validation curves for these models are included in Appendix B.

Other behavioral and attitudinal variables exhibiting a negative relationship with consumption have relatively small standardized coefficients.

Of these selected variables, nine measure characteristics of the dwelling and appliance or energy-related device ownership. Seven measure occupant socio-demographics, six measure attitudinal factors, and six measure behavioral factors. For the top ten variables with the strongest associations to electricity consumption, seven are either dwelling characteristics or socio-demographics, two are behavioral variables, and one is an attitude variable.

6. Discussion

6.1. Summary of results and comparison to previous research

The methodological results of this study show that the regularization methods introduced in this paper achieve a RMSE on the test set ranging from 0.4562–0.4457, which is equivalent to 0.81–0.79 of the response variable's standard deviation. In other words, the prediction error of these methods is 20% smaller than the standard deviation of log-transformed annual electricity consumption. These prediction error results compare favorably with those of other studies employing regularization methods for building energy consumption prediction [10, 36, 37]. Across methods, the other goodness-of-fit measure, R-squared (ranging from 0.375–0.396) is consistent with or surpasses results of many previous studies of household electricity usage [71, 72, 22, 49, 122, 47].

Returning to the question of comparing model predictive accuracy with sparsity, the regularization methods (excluding ridge regression) yield a sizable reduction in the number of variables needed to achieve similar predictive accuracy. Comparing adaptive lasso and stepwise regression, for instance, adaptive lasso selects a model that has less than a 2% greater prediction error than stepwise regression but reduces the number of variables in the model by a further 27%. Trading off a small increase in prediction error for a large reduction in the number of variables that needs to be collected is favorable when the objective of analysis includes a simpler, more interpretable model.

Predictor	$\hat{\beta}_{elastic}$	$\hat{\beta}_{gLasso}$	$\hat{\beta}_{adLasso}$
High usage predictors			
EV ownership	0.155	0.102	0.214
Size of home	0.115	0.136	0.138
Occupancy level	0.080	0.109	0.089
Type of dwelling	0.098	0.036	0.109
Important value: comfort	0.053	0	0.055
Household income	0.039	0.063	0
Solar water heating	0.028	0	0.064
Time-of-use rate	0.023	0	0.057
AC ownership	0.025	0.021	0.018
Age	0.002	0.041	0
Roof insulation	0.013	0.027	0
Number of bedrooms	0.023	0.016	0
Other device ownership	0	0	0.023
Renewables too expensive	0	0.015	0
Talk with family about bill	0.001	0.011	0
Smart thermostat ownership	0.011	0	0
Gender	0.010	0	0
Talk with family about conservation	0	0.003	0
Low usage predictors			
Behavior: turn off AC	−0.094	0	−0.186
Behavior: unplug appliances	−0.075	−0.089	−0.062
Rents home	−0.071	0	−0.077
Behavior: take a shorter shower	−0.006	−0.023	0
Important value: cost of energy	−0.019	0	0
New technologies will solve problems	0	−0.018	0
Benefit of renewables	−0.010	0	0
Behavior: turn off lights	−0.008	0	0
Would do more to save if I knew how	0	−0.001	0

Table 9: Variables selected across elastic net, group lasso, and adaptive lasso models. Variables are split by the sign of their effect and are ordered by the magnitude of their standardized coefficient.

The empirical results of this study confirm that the size of home and number of occupants are two of the strongest determinants of residential electricity use patterns [43, 46, 22, 48, 45].

Type of dwelling [55, 45, 22] and income [53, 55, 46] can be confirmed as strong predictors even though results from other studies are mixed on their effect [e.g. 44, 48]. Previous findings on the associations between tenure type and electricity consumption are similarly mixed, with some supporting this study's findings of higher consumption in privately-owned residences [47, 45, 55], while others report either higher consumption in rented buildings or no significant effect [44, 61, 48, 32].

Unsurprisingly, EV ownership and presence of AC in the home both exhibit positive associations with annual electricity use. Given the growing uptake of these technologies around the world, a detailed understanding of their impact on total consumption is increasingly important.

Occupant attitudes toward energy conservation and renewable energy, the values occupants consider important in relation to energy use, and their knowledge of energy concepts do not exhibit strong associations with electricity consumption. These results support those of previous studies that do not find a notable link between environmental attitudes and electricity consumption [75, 74, 76]. Furthermore, rates of energy knowledge as demonstrated by performance on an energy quiz bear little association to electricity usage, which is similar to the findings of Brounen et al. [76]. One attitude variable is particularly strong in comparison to other predictors: listing 'comfort' as one of the most important values related to energy use is associated with higher consumption. This supports Wilhite et al.'s [123] argument that notions of comfort and convenience may have considerable implications for electricity demand and are not sufficiently addressed in energy demand research.

From the set of behavior variables, unplugging appliances when not in use for extended periods and turning off AC when not needed are selected as predictors of lower usage in the model. These results confirm those of Wallis et al. [84], who find a statistically significant association between habitual energy saving behaviors and reduced annual consumption.

Participation in energy efficiency programs and uptake of rebates for efficient appliances are not among the significant predictors in the model. This may be due to very low rates of participation and uptake reported amongst the survey sample.

6.2. Implications of results

These results have implications both for statistical approaches for modeling building electricity consumption as well as for understanding factors that influence consumption.

Given the complexities of residential electricity consumption, statistical methods that reduce large predictor sets without sacrificing much predictive accuracy are advantageous in studies of domestic electricity demand. The regularization methods introduced in this paper, including extensions to the lasso that take into consideration some of its methodological weaknesses, are useful in this regard. Furthermore, these methods are computationally efficient and can address several important statistical challenges, such as model overfitting, multicollinearity, and high-dimensional data. Even in the absence of these issues, the methods presented here can effectively identify key variables in models of building energy consumption, and they do not suffer from the same statistical weaknesses as do other variable selection approaches. For these reasons, they are especially suitable for building energy modeling, give the specific challenges faced in this discipline.

This paper has stressed the importance of letting the analysis objectives and the characteristics of the data guide the use of regularization methods. It has explained why, for instance, both elastic net and lasso are likely to show similar results given the absence of strong multicollinearity effects and high-dimensionality, and it has confirmed this empirically (lasso and elastic net select the same variables, although prediction error is somewhat higher for lasso). The extensions to the lasso are introduced to improve upon these results, and we see that they do (in terms of yielding simpler models without much loss in predictive accuracy).

The empirical implications of this study are best understood in the context of the study location. Palo Alto's population is projected to grow at a rate of 1.1% annually over the next 20 years. The city's senior population (65 and over) is one of its fastest growing demographics [124]. In this region, large, detached homes are commonplace, occupancy levels are growing, and the city's average median family income is the third highest in the U.S. [125]. Given the demonstrated effects of dwelling size and type, occupancy levels, and household income on residential electricity consumption, these trends are important to consider when determining ways to meet the city's ambitious energy savings targets.

This study provides evidence that policies or programs that further improve the thermal performance and efficiency of residential buildings are necessary to achieve substantial emissions reductions. This evidence may be especially relevant for single-family, detached homes in Palo Alto. Given the study's findings that energy efficiency program uptake is low, more effort is needed to engage residential customers in this regard. Encouraging regular home energy audits through building codes and regulations could help determine where home efficiencies are lacking. A target audience for these initiatives should be the city's older residential population, as a link was found between age and electricity consumption in the models.

Despite Palo Alto's relatively mild climate (less than half the sample owned AC), the significance of AC for consumption suggests that reducing AC use deserves special attention. This is even more pressing given the anticipated rise in home AC ownership in middle-income countries with additional warming. Davis and Gertler [126] predict near-universal saturation of AC in all warm areas in just a few decades, and their findings suggest AC impacts on energy usage will be larger than previously believed. AC adoption and use must be met with even greater energy efficiency gains or behavioral changes to reduce its projected impact, especially considering the effect of AC use on peak demand.

The same applies to EV ownership. Palo Alto has one of the highest rates of EV ownership in the country (around 3–4% of registered vehicles) and aims for 90% of registered vehicles to be electric by 2030. This study provides further evidence that EV ownership must be met with vehicle-to-grid integration projects and smart charging policies to lessen the substantial burden this transformation will place on local electricity networks [127].

This study provides evidence that households can decrease their electricity usage by engaging more frequently in energy saving behaviors, especially those related to appliances and AC. While 70% of respondents report they 'Always' turn off lights and appliances when not in use, only 20% report the same for unplugging their appliances. Further savings could be achieved given that standby power consumption is responsible for around 15% of household electricity usage in California [128]. Much of the focus in reducing residential electricity consumption has been on deploying energy efficiency measures rather than motivating changes in behavior, but this study highlights the important role of habitual actions taken to save energy in the home and reaffirms previous findings that these can contribute towards reducing carbon emissions [129].

6.3. Limitations

This study has limitations to its design, methods, and data. In terms of its design, the sampling methodology is non-random, and participation is limited to utility customers with emails on record. Some of the biases, such as underrepresentation of renters and people in the 18–35 age group, have been discussed. This means that the findings are not necessarily generalizable. The self-reporting of behaviors and attitudes could mean social desirability bias is present and may have influenced results [130].

The regularization methods applied in this study show promise for improving building energy prediction and selecting sparse models that highlight key variables. Their application in this study, however, does not showcase their suitability for addressing other issues, such as multicollinearity and overfitting, since these challenges are muted in the data. The cross-validation results suggest the models do not exhibit high variance, even without applying much regularization. This is likely because the sample size is large compared to the number of predictors. With a smaller sample size, or with an increasingly large number of predictors, the regularization methods introduced here are likely to improve in performance, especially in their prediction error on the test set. Some evidence of this is seen when fitting the models to the data including only those households with AC ($N = 390$). Cross-validation curves for these data are slightly more U-shaped (see Appendix B). One further methodological limitation is that the analysis did not consider interactions between predictors. Evidence from previous research suggests these methods and several extensions can handle high numbers of pair-wise interactions, which could enable further insight into the drivers of building energy consumption [10].

Regarding the limitations to the data, additional details on appliance ownership and use may increase the explanatory power of the models and yield deeper insights into how occupant behavior is associated with electricity consumption. Other specific factors not investigated include more detailed efficiency measures taken in the home, data on the type of AC (central versus window unit), pool ownership, and fuel used for space heating. The last of these may be particularly important, given an estimated 25% of Palo Alto households use electricity for heating [116]. Nine respondents indicated ownership of air source heat pumps on the 'Other' device survey question, but a specific question on fuel used for heating could have revealed the influence of electric heating on annual consumption.

Furthermore, this paper is limited in explaining the drivers of specific electricity end-uses, such as space heating and cooling, water heating, or appliances, lighting and electronics, which makes comparing results to other study contexts more difficult [43]. Similarly, the analyses presented here only consider electricity and not natural gas consumption. The modeling techniques presented could be applied to natural gas usage data for further insight on how to reduce residential building emissions from space heating and cooking.

7. Conclusions

This paper discusses the use of regularization methods in linear regression analysis for improving both prediction and interpretation in residential building energy models. It identifies key challenges in energy modeling and explains how regularization methods can address these. Next, it demonstrates these methods empirically on multivariate survey and household electricity data for a sample of 1,008 households in Palo Alto, California. It tests a wide range of structural and occupant factors across several distinct variable types to determine those exhibiting the strongest associations with annual electricity use.

The results show that regularization methods can improve upon traditional variable selection approaches, such as stepwise regression, both in terms of prediction error and model interpretability. Elastic net and group lasso make better predictions on hold-out test data than the other methods while reducing the number of nonzero coefficients in the models. Adaptive lasso selects the most sparse model with 11 predictors, a reduction of over 80%, with only a 1-2% higher prediction error than the other methods.

The analysis finds that household electricity use is best explained through a combination of socio-demographic and physical dwelling characteristics. Size of home, occupancy levels, and ownership of an EV and AC are significantly associated with increased electricity usage. While occupants' attitudes toward the environment and their level of energy knowledge do not generally show strong associations with consumption, this paper does find that specific occupant curtailment behaviors, such as unplugging appliances when not in use for extended periods and turning off AC when no one is home, are strong predictors of lower electricity use.

These findings can inform Palo Alto's energy strategy as it embarks on ambitious usage reduction targets over the next several decades. Results are also informative for other cities and regions that want to understand the key variables influencing consumption, or want to use these to better predict future patterns of consumption.

While the evidence presented here does not refute the importance of improving the structural efficiency of the building stock in order to achieve these targets, it also presents evidence that occupant factors related to curtailment behavior are drivers of electricity consumption. This insight is particularly important for designing energy policy in places that expect rapid increases in EV and AC ownership. In Palo Alto, these are expected to be near-universal in the California Bay Area by 2050 [113]. Reducing home size and occupancy levels are more challenging policy changes to implement than are encouraging energy curtailment behaviors. Of course, understanding the most effective ways to do this is of equal importance and is the subject of much ongoing research. Here, especially, is where other disciplinary approaches that examine the socio-technical structures surrounding behaviors or practices may add the most insight.

Situating these results within a growing body of research on the factors that drive household electricity consumption will contribute to future lines of empirical inquiry in this field. The purpose of future research should be to further investigate the links between the factors identified and tested in this paper as well as to explore any number of additional influential factors that influence household electricity consumption. The methods demonstrated in this paper are applicable to a wide variety of energy and building data, and they can be used successfully in contexts where other statistical methods fail (e.g. where the issues of multicollinearity and high-dimensionality are present). Of particular interest for further research is the application of these methods to higher-resolution electricity data. Drivers of electricity consumption across months and years may be different than those that influence daily or hourly consumption patterns. Understanding these differences through the use of regularization in statistical models can inform strategies for reducing demand on both of these time-scales, which is increasingly important for a low-carbon transition.

Acknowledgements

The author would like to thank the editor and several anonymous reviewers for their detailed comments, which helped clarify the structure and aim of the paper.

Thanks also to the City of Palo Alto Utilities for providing usage data and for facilitating deployment of the survey. Thanks especially to Lacey Lutes and Lisa Benatar for their assistance. Thank you to Phil Grünewald, John Fresen, and Pete Walton for their instructive comments and feedback. Any remaining errors in this article are solely the work of the author.

The author gratefully acknowledges financial support for this research from the City of Palo Alto Utilities, the University of Oxford's Environmental Change Institute, and the Sir Peter Elworthy fund. The author declares no competing interests.

Appendix A.

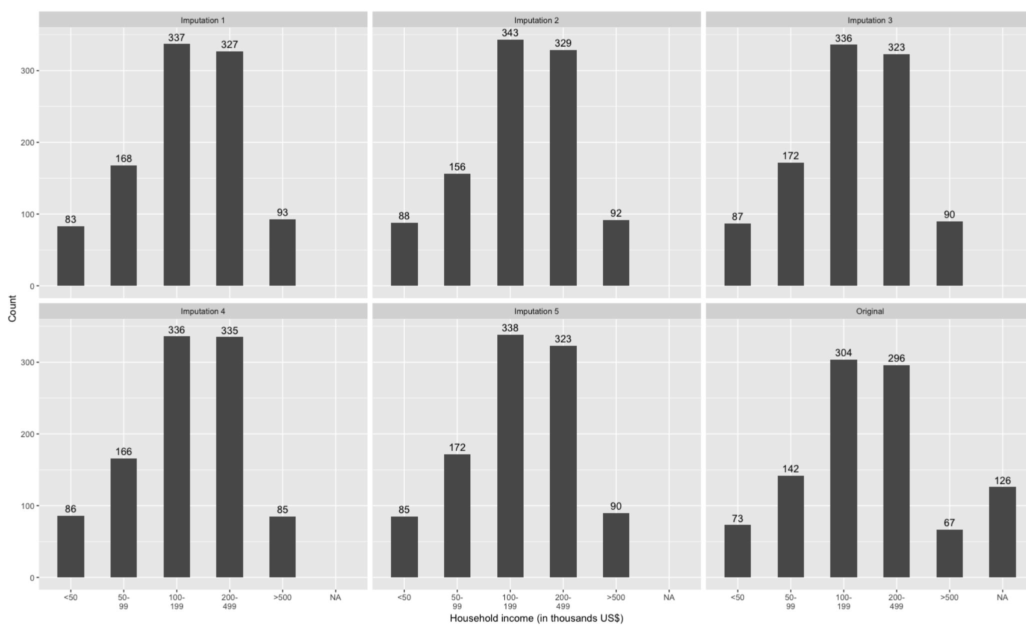

Figure A.4: Distributions for *Income* variable across five imputed datasets compared with original distribution.

Appendix B.

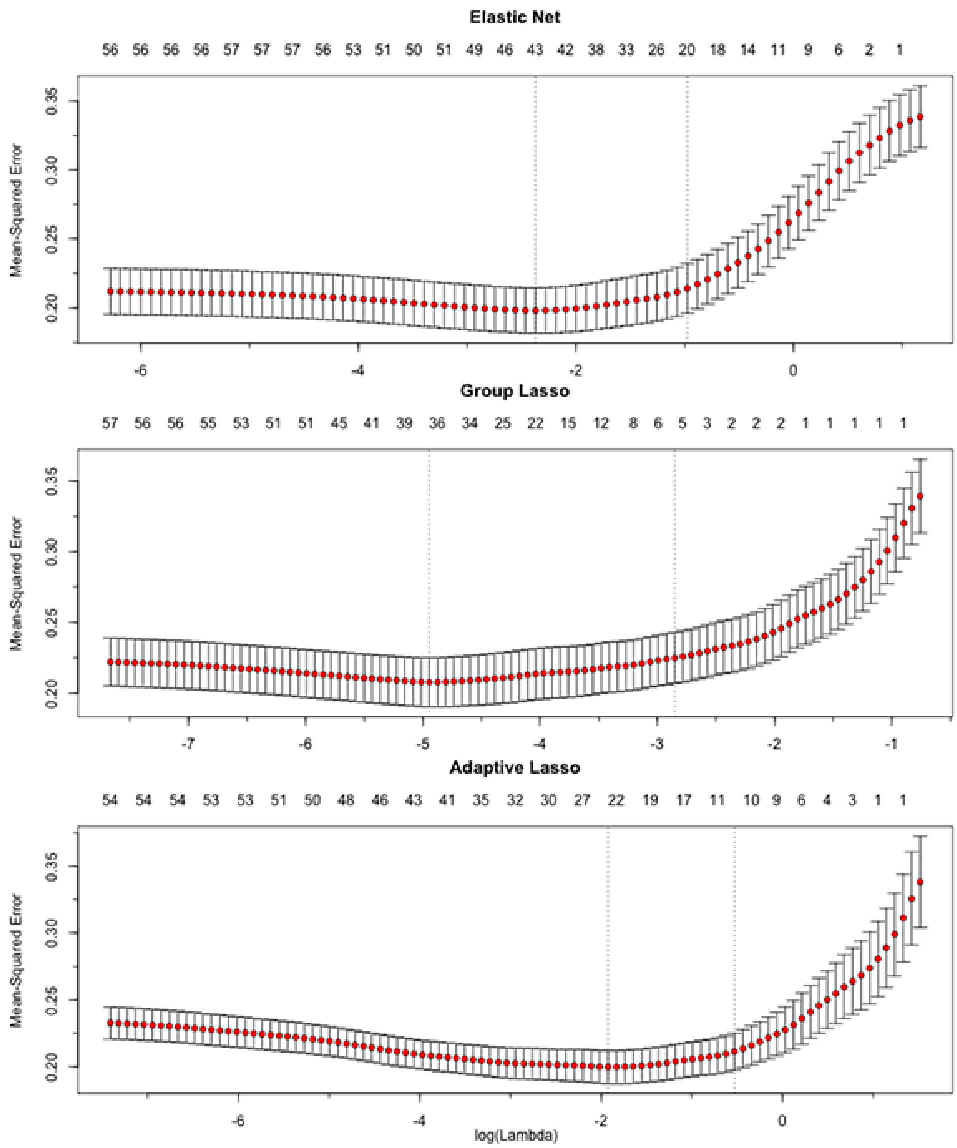

Figure B.5: Plots of cross-validation MSE for elastic net, group lasso, and adaptive lasso applied to the data filtered for AC ownership ($N = 390$). Axes and plots elements are the same as in Figures 2–3.

References

[1] P. Nejat, F. Jomehzadeh, M. M. Taheri, M. Gohari, M. Z. Abd. Majid, A global review of energy consumption, CO2 emissions and policy in the residential sector (with an overview of the top ten CO2 emitting countries), Renewable and Sustainable Energy Reviews 43 (2015) 843–862.

[2] EPA, Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2015, Technical Report, U.S. Environmental Protection Agency, 2017.

[3] U.S. EIA, Annual Energy Outlook 2018 with Projections to 2050, Technical Report, U.S. Energy Information Administration, Washington, D.C., 2018.

[4] J. L. Reyna, M. V. Chester, Energy efficiency to reduce residential electricity and natural gas use under climate change, Nature Communications 8 (2017) 14916.

[5] EPA, Regulatory Impact Analysis for the Clean Power Plan Final Rule, Technical Report EPA-452/R-15-003, U.S. Environmental Protection Agency, Research Triangle Park, NC, 2015.

[6] I. M. Hoffman, C. A. Goldman, S. Murphy, N. A. Frick, G. Leventis, L. C. Schwartz, The Cost of Saving Electricity Through Energy Efficiency Programs Funded by Utility Customers: 2009–2015, Technical Report 1457014, Lawrence Berkeley National Laboratory, 2018.

[7] A. Grubler, C. Wilson, N. Bento, B. Boza-Kiss, V. Krey, D. L. McCollum, N. D. Rao, K. Riahi, J. Rogelj, S. De Stercke, J. Cullen, S. Frank, O. Fricko, F. Guo, M. Gidden, P. Havlík, D. Huppmann, G. Kiesewetter, P. Rafaj, W. Schoepp, H. Valin, A low energy demand scenario for meeting the 1.5 °C target and sustainable development goals without negative emission technologies, Nature Energy 3 (2018) 515–527.

[8] L. Mundaca, D. Ürge-Vorsatz, C. Wilson, Demand-side approaches for limiting global warming to 1.5 °C, Energy Efficiency (2018).

[9] K. Amasyali, N. M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews 81 (2018) 1192–1205.

[10] D. Hsu, Identifying key variables and interactions in statistical models of building energy consumption using regularization, Energy 83 (2015) 144–155.

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

[12] M. Pampaka, G. Hutcheson, J. Williams, Handling missing data: Analysis of a challenging data set using multiple imputation, International Journal of Research & Method in Education 39 (2016) 19–37.

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

[14] H.-X. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renewable and Sustainable Energy Reviews 16 (2012) 3586–3592.

[15] H.-X. Zhao, F. Magoulès, Parallel Support Vector Machines Applied to the Prediction of Multiple Buildings Energy Consumption, Journal of Algorithms & Computational Technology 4 (2010) 231–249.

[16] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renewable and Sustainable Energy Reviews 82 (2018) 1027–1047.

[17] Z. Yu, F. Haghighat, B. C. M. Fung, H. Yoshino, A decision tree method for building energy demand modeling, Energy and Buildings 42 (2010) 1637–1646.

[18] Z. Wang, Y. Wang, R. Zeng, R. S. Srinivasan, S. Ahrentzen, Random Forest based hourly building energy prediction, Energy and Buildings 171 (2018) 11–25.

[19] G. Shmueli, To Explain or to Predict?, Statistical Science 25 (2010) 289–310.

[20] D. E. Farrar, R. R. Glauber, Multicollinearity in Regression Analysis: The Problem Revisited, The Review of Economics and Statistics 49 (1967) 92.

[21] N. Fumo, M. Rafe Biswas, Regression analysis for prediction of residential energy consumption, Renewable and Sustainable Energy Reviews 47 (2015) 332–343.

[22] G. Huebner, D. Shipworth, I. Hamilton, Z. Chalabi, T. Oreszczyn, Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes, Applied Energy 177 (2016) 692–702.

[23] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, volume 1 of *Springer Series in Statistics*, Springer, Berlin, 2001.

[24] H. T. Pao, Forecasting energy consumption in Taiwan using hybrid nonlinear models, Energy 34 (2009) 1438–1446.

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

[26] G. Heinze, C. Wallisch, D. Dunkler, Variable selection – A review and recommendations for the practicing statistician, Biometrical Journal 60 (2018) 431–449.

[27] C. Filippín, F. Ricard, S. Flores Larsen, Evaluation of heating energy consumption patterns in the residential building sector using stepwise selection and multivariate analysis, Energy and Buildings 66 (2013) 571–581.

[28] A. Kialashaki, J. R. Reisel, Modeling of the energy demand of the residential sector in the United States using regression models and artificial neural networks, Applied Energy 108 (2013) 271–280.

[29] M. Wang, J. Wright, A. Brownlee, R. Buswell, A comparison of approaches to stepwise regression on variables sensitivities in building simulation and analysis, Energy and Buildings 127 (2016) 313–326.

[30] C. Deb, S. E. Lee, Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data, Energy and Buildings 159 (2018) 228–245.

[31] J. C. Lam, K. K. W. Wan, K. L. Cheung, L. Yang, Principal component analysis of electricity use in office buildings, Energy and Buildings 40 (2008) 828–836.

[32] D. Ndiaye, K. Gabriel, Principal component analysis of the electricity consumption in residential dwellings, Energy and Buildings 43 (2011) 446–453.

[33] N. Djuric, V. Novakovic, Identifying important variables of energy use in low energy office building by using multivariate analysis, Energy and Buildings 45 (2012) 91–98.

[34] F. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer Series in Statistics, Springer-Verlag, New York, 2001.

[35] B. Ratner, Variable selection methods in regression: Ignorable problem, outing notable solution, Journal of Targeting, Measurement and Analysis for Marketing 18 (2010) 65–75.

[36] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M. A. Brown, R. M. Pendyala, Machine learning approaches for estimating commercial building energy consumption, Applied Energy 208 (2017) 889–904.

[37] H. Deng, D. Fannon, M. J. Eckelman, Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata, Energy and Buildings 163 (2018) 34–43.

[38] Jain R. K., Damoulas T., Kontokosta C. E., Towards Data-Driven Energy Consumption Forecasting of Multi-Family Residential Buildings: Feature Selection via The Lasso, in: 2014 International Conference

on Computing in Civil and Building Engineering, American Society of Civil Engineers, Orlando, Florida, 2014, pp. 1675–1682.

[39] B. Uniejewski, J. Nowotarski, R. Weron, Automated Variable Selection and Shrinkage for Day-Ahead Electricity Price Forecasting, Energies 9 (2016) 621.

[40] Y. Guo, J. Wang, H. Chen, G. Li, J. Liu, C. Xu, R. Huang, Y. Huang, Machine learning-based thermal response time ahead energy demand prediction for building heating systems, Applied Energy 221 (2018) 16–27.

[41] G. Suryanarayana, J. Lago, D. Geysen, P. Aleksiejuk, C. Johansson, Thermal load forecasting in district heating networks using deep learning and advanced feature selection methods, Energy 157 (2018) 141–149.

[42] F. Belaïd, D. Roubaud, E. Galariotis, Features of residential energy consumption: Evidence from France using an innovative multilevel modelling approach, Energy Policy 125 (2019) 277–285.

[43] R. V. Jones, A. Fuertes, K. J. Lomas, The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings, Renewable and Sustainable Energy Reviews 43 (2015) 901–917.

[44] M. Bedir, E. Hasselaar, L. Itard, Determinants of electricity consumption in Dutch dwellings, Energy and Buildings 58 (2013) 194–207.

[45] P. Wyatt, A dwelling-level investigation into the physical and socio-economic drivers of domestic energy consumption in England, Energy Policy 60 (2013) 540–549.

[46] D. Brounen, N. Kok, J. M. Quigley, Residential energy use and conservation: Economics and demographics, European Economic Review 56 (2012) 931–945.

[47] D. Wiesmann, I. Lima Azevedo, P. Ferrão, J. E. Fernández, Residential electricity consumption in Portugal: Findings from top-down and bottom-up models, Energy Policy 39 (2011) 2772–2779.

[48] A. Kavousian, R. Rajagopal, M. Fischer, Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior, Energy 55 (2013) 184–194.

[49] F. McLoughlin, A. Duffy, M. Conlon, Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study, Energy and Buildings 48 (2012) 240–248.

[50] O. Guerra-Santin, L. Itard, Occupants' behaviour: Determinants and effects on residential heating consumption, Building Research & Information 38 (2010) 318–338.

[51] P. Tiwari, Architectural, demographic, and economic causes of electricity consumption in Bombay, Journal of Policy Modeling 22 (2000) 81–98.

[52] J.-M. Cayla, N. Maizi, C. Marchand, The role of income in energy consumption behaviour: Evidence from French households data, Energy Policy 39 (2011) 7874–7883.

[53] J. C. Cramer, N. Miller, P. Craig, B. M. Hackett, T. M. Dietz, E. L. Vine, M. D. Levine, D. J. Kowalczyk, Social and engineering determinants and their equity implications in residential electricity use, Energy 10 (1985) 1283–1291.

[54] S. Kelly, Do homes that are more energy efficient consume less energy?: A structural equation model of the English residential sector, Energy 36 (2011) 5610–5620.

[55] Y. G. Yohanis, J. D. Mondol, A. Wright, B. Norton, Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use, Energy and Buildings 40 (2008) 1053–1059.

[56] K. Gram-Hanssen, Efficient technologies or user behaviour, which is the more important when reducing households' energy consumption?, Energy Efficiency 6 (2013) 447–457.

[57] A.-L. Lindén, A. Carlsson-Kanyama, B. Eriksson, Efficient and inefficient aspects of residential energy behaviour: What are the policy instruments for change?, Energy Policy 34 (2006) 1918–1927.

[58] V. Assimakopoulos, Residential energy demand modelling in developing regions: The use of multivariate statistical techniques, Energy Economics 14 (1992) 57–63.

[59] E. Hirst, R. Goeltz, Comparison of Actual Energy Savings with Audit Predictions for Homes in the North Central Region of the U.S.A., Building and Environment 20 (1985) 1–6.

[60] G. K. Tso, J. Guan, A multilevel regression approach to understand effects of environment indicators and household features on residential energy consumption, Energy 66 (2014) 722–731.

[61] G. K. Tso, K. K. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy 32 (2007) 1761–1768.

[62] T. F. Sanquist, H. Orr, B. Shui, A. C. Bittner, Lifestyle factors in U.S. residential electricity consumption, Energy Policy 42 (2012) 354–364.

[63] S. Darby, The Effectiveness of Feedback on Energy Consumption, Technical Report, Environmental Change Institute, University of Oxford, 2006.

[64] W. Abrahamse, L. Steg, C. Vlek, T. Rothengatter, The effect of tailored information, goal setting, and tailored feedback on household energy use, energy-related behaviors, and behavioral antecedents, Journal of Environmental Psychology 27 (2007) 265–276.

[65] C. Fischer, Feedback on household electricity consumption: A tool for saving energy?, Energy Efficiency 1 (2008) 79–104.

[66] P. W. Schultz, M. Estrada, J. Schmitt, R. Sokoloski, N. Silva-Send, Using in-home displays to provide smart meter feedback about household electricity consumption: A randomized control trial comparing kilowatts, cost, and social norms, Energy 90 (2015) 351–358.

[67] J. H. Van Houwelingen, W. F. Van Raaij, The effect of goal-setting and daily electronic feedback on in-home energy use, Journal of Consumer Research 16 (1989) 98–105.

[68] K. Ehrhardt-Martinez, K. A. Donnelly, S. Laitner, Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities, Technical Report, ACEEE, 2010.

[69] S. Huang, H. Safiullah, J. Xiao, B.-M. S. Hodge, R. Hoffman, J. Soller, D. Jones, D. Dininger, W. E. Tyner, A. Liu, J. F. Pekny, The effects of electric vehicles on residential households in the city of Indianapolis, Energy Policy 49 (2012) 442–455.

[70] DOE, Evaluating Electric Vehicle Charging Impacts and Customer Charging Behaviors - Experiences from Six Smart Grid Investment Grant Projects, Technical Report, U.S. Department of Energy Office of Electricity Delivery & Energy Reliability, Washington, D.C., 2014.

[71] W. Abrahamse, L. Steg, How do socio-demographic and psychological factors relate to households' direct and indirect energy use and savings?, Journal of Economic Psychology 30 (2009) 711–720.

[72] G. Brandon, A. Lewis, Reducing household energy consumption: A qualitative and quantitative field study, Journal of Environmental Psychology 19 (1999) 75–85.

[73] D. Gadenne, B. Sharma, D. Kerr, T. Smith, The influence of consumers' environmental beliefs and attitudes on energy saving behaviours, Energy Policy 39 (2011) 7684–7694.

[74] K. Vringer, T. Aalbers, K. Blok, Household energy requirement and value patterns, Energy Policy 35 (2007) 553–566.

[75] F. Bartiaux, K. Gram-Hanssen, Socio-political factors influencing household electricity consumption: A comparison between Denmark and Belgium, in: Energy Savings: What Works & Who Delivers, volume 3, European Council for an Energy Efficient Economy, Mandelieu la Napoule, France, 2005, pp. 1313–1325.

[76] D. Brounen, N. Kok, J. M. Quigley, Energy literacy, awareness, and conservation behavior of residential households, Energy Economics 38 (2013) 42–50.

[77] NEETF & RoperASW, Americans' Low Energy IQ:" A Risk to Our Energy Future. Why America Needs a Refresher Course on Energy, Technical Report, The National Environmental Education & Training Foundation & Roper ASW, 2002.

[78] K. Gram-Hanssen, New needs for better understanding of household's energy consumption – behaviour, lifestyle or practices?, Architectural Engineering and Design Management 10 (2014) 91–107.

[79] I. Mansouri, M. Newborough, D. Probert, Energy Consumption in UK Households: Impact of Domestic Electrical Appliances, Applied Energy 54 (1996) 211–285.

[80] S. Barr, A. W. Gilg, N. Ford, The household energy gap: Examining the divide between habitual- and purchase-related conservation behaviours, Energy Policy 33 (2005) 1425–1444.

[81] G. T. Gardner, P. C. Stern, The Short List: The Most Effective Actions U.S. Households Can Take to Curb Climate Change, Environment Magazine (2009).

[82] K. Steemers, G. Y. Yun, Household energy consumption: A study of the role of occupants, Building Research & Information 37 (2009) 625–637.

[83] H. Brohus, P. K. Heiselberg, A. Simonsen, K. C. Sørensen, Influence of Occupants' Behaviour on the Energy Consumption of Domestic Buildings, in: Clima 2010: 10th Rehva World Congress: Sustainable Energy Use in Buildings, Wiley series in probability and statistics, Clima 2010: 10th Rehva World Congress, Antalya, 2010.

[84] H. Wallis, M. Nachreiner, E. Matthies, Adolescents and electricity consumption; Investigating sociodemographic, economic, and behavioural influences on electricity consumption in households, Energy Policy 94 (2016) 224–234.

[85] J. Thøgersen, A. Grønhøj, Electricity saving in households—A social cognitive approach, Energy Policy 38 (2010) 7732–7743.

[86] G. M. Huebner, J. Cooper, K. Jones, Domestic energy consumption—What role do comfort, habit, and knowledge about the heating system play?, Energy and Buildings 66 (2013) 626–636.

[87] S. Sorrell, J. Dimitropoulos, M. Sommerville, Empirical estimates of the direct rebound effect: A review, Energy Policy 37 (2009) 1356–1371.

[88] E. H. Kennedy, T. M. Beckley, B. L. McFarlane, S. Nadeau, Why we don't walk the talk': Understanding the environmental values/behaviour gap in Canada, Human Ecology Review 16 (2009) 151.

[89] A. E. Hoerl, R. W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics 12 (1970) 55.

[90] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.

[91] H. Zou, T. Hastie, Regularization and Variable Selection via the Elastic Net, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 67 (2005) 301–320.

[92] N. Roberts, J. Thatcher, Conceptualizing and testing formative constructs: Tutorial and annotated example, ACM SIGMIS Database 40 (2009) 9–39.

[93] A. Diamantopoulos, The error term in formative measurement models: Interpretation and modeling implications, Journal of Modelling in Management 1 (2006) 7–17.

[94] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (2006) 49–67.

[95] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the Lasso, The Annals of Statistics 34 (2006) 1436–1462.

[96] P. Zhao, B. Yu, On Model Selection Consistency of Lasso, Journal of Machine Learning Resesearch 7 (2006) 2541–2563.

[97] J. Fan, R. Li, Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, Journal of the American Statistical Association 96 (2001) 1348–1360.

[98] H. Zou, The Adaptive Lasso and Its Oracle Properties, Journal of the American Statistical Association 101 (2006) 1418–1429.

[99] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, Journal of Statistical Software 33 (2010) 1.

[100] J. Taylor, R. J. Tibshirani, Statistical learning and selective inference, Proceedings of the National Academy of Sciences 112 (2015) 7629–7634.

[101] R. Lockhart, J. Taylor, R. J. Tibshirani, R. Tibshirani, A significance test for the lasso, Annals of statistics 42 (2014) 413.

[102] J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor, Exact post-selection inference, with application to the lasso, The Annals of Statistics 44 (2016) 907–927.

[103] G. King, J. Honaker, A. Joseph, K. Scheve, Analyzing incomplete political science data: An alternative algorithm for multiple imputation, American Political Science Review 95 (2001) 49–69.

[104] J. Honaker, G. King, What to do about missing values in time-series cross-section data, American Journal of Political Science 54 (2010) 561–581.

[105] D. B. Rubin, Multiple Imputation After 18+ Years, Journal of the American Statistical Association 91 (1996) 473.

[106] J. Honaker, G. King, M. Blackwell, Amelia II: A Program for Missing Data, 2015.

[107] M. Takahashi, Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations, Data Science Journal 16 (2017) 37.

[108] R Core Team, R: A language and environment for statistical computing., R Foundation for Statistical Computing, 2016.

[109] Y. Yang, H. Zou, Group Lasso Penalized Learning Using a Unified BMD Algorithm, 2017.

[110] K. Max, Caret: Classification and Regression Training, 2018.

[111] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, D. Firth, MASS: Support Functions and Datasets for Venables and Ripley's MASS, 2018.

[112] WRCC, Palo Alto, California: Period of Record General Climate Summary 1953-2006, https://wrcc.dri.edu/cgi-bin/cliMAIN.pl?capalo+sfo, 2006.

[113] City of Palo Alto, Palo Alto Sustainability and Climate Action Plan: Framework, Principles, Guidelines, Goals & Strategies, Technical Report, City of Palo Alto, Palo Alto, California, 2016.

[114] CEC, Total Electric System Generation, http://www.energy.ca.gov/almanac/electricity_data/total_system_power.html, 2017.

[115] EIA, How much electricity does an American home use?, https://www.eia.gov/tools/faqs/faq.php?id=97&t=3, 2016.

[116] U.S. Census Bureau, 2011-2015 American Community Survey, https://www.census.gov/acs/www/data/data-tables-and-tools/, 2015.

[117] J. DeWaters, S. Powers, Establishing Measurement Criteria for an Energy Literacy Questionnaire, The Journal of Environmental Education 44 (2013) 38–55.

[118] J. DeWaters, B. Qaqish, M. Graham, S. Powers, Designing an Energy Literacy Questionnaire for Middle and High School Youth, The Journal of Environmental Education 44 (2013) 56–78.

[119] K. Coyle, Environmental literacy in America: What ten years of NEETF/Roper research and related studies say about environmental literacy in the US, National Environmental Education & Training Foundation (2005).

[120] B. Southwell, J. Murphy, J. DeWaters, P. LeBaron, Americans' Perceived and Actual Understanding of Energy, Technical Report, RTI Press, Research Triangle Park, NC, 2012.

[121] R. Halvorsen, R. Palmquist, The Interpretation of Dummy Variables in Semilogarithmic Equations, The American Economic Review 70 (1980) 474–475.

[122] M. H. Wahlström, B. Hårsman, Residential energy consumption and conservation, Energy and Buildings 102 (2015) 58–66.

[123] H. Wilhite, E. Shove, L. Lutzenhiser, W. Kempton, Twenty years of energy demand management: We know more about individual behavior but how much do we really know about demand, in: Proceedings of the 2000 ACEEE Summer Study on Energy Efficiency in Buildings, pp. 435–453.

[124] ABAG, San Francisco Bay Area State of the Region: Economy, Population, Housing 2015, Technical Report, Association of Bay Area Governments, Oakland, CA, 2015.

[125] J. Pisillo, Palo Alto residents earn 3rd highest median family income, http://blog.sfgate.com/ontheblock/2012/10/12/palo-alto-residents-earn-3rd-highest-median-family-income/, 2012.

[126] L. W. Davis, P. J. Gertler, Contribution of air conditioning adoption to future energy use under global warming, Proceedings of the National Academy of Sciences 112 (2015) 5962–5967.

[127] I. Sidhu, P. Kaminsky, B. Tenderich, N. DeForest, A. Lorimer, B. Ur, Impact of Widespread Electric Vehicle Adoption on the Electrical Utility Business–Threats and Opportunities, Technical Report, University of California, Berkeley Center for Entrepreneurship & Technology (CET), 2009.

[128] LBNL, Developing and Testing Low Power Mode Measurement Methods, Technical Report P500-04-057, Prepared for California Energy Commission Public Interest Energy Research Programme by Lawrence Berkeley National Laboratory, 2004.

[129] T. Dietz, G. T. Gardner, J. Gilligan, P. C. Stern, M. P. Vandenbergh, Household actions can provide a behavioral wedge to rapidly reduce US carbon emissions, Proceedings of the National Academy of Sciences 106 (2009) 18452–18456.

[130] R. J. Fisher, Social desirability bias and the validity of indirect questioning, Journal of Consumer Research 20 (1993) 303–315.