Journal of Global Information Management
Volume 29 • Issue 2 • Bi-Monthly 2021

Modeling Residential Energy Consumption:
An Application of IT-Based Solutions and Big Data Analytics for Sustainability

Roya Gholami, NEOMA Business School, France
Rohit Nishant, Laval University, Canada
Ali Emrouznejad, Aston University, UK
https://orcid.org/0000-0001-8094-4244

ABSTRACT

Smart meters that allow information to flow between users and utility service providers are expected to foster intelligent energy consumption. Previous studies focusing on demand-side management have been predominantly restricted to factors that utilities can manage and manipulate, but have ignored factors specific to residential characteristics. They also often presume that households consume similar amounts of energy and electricity. To fill these gaps in literature, the authors investigate two research questions: (RQ1) Does a data mining approach outperform traditional statistical approaches for modelling residential energy consumption? (RQ2) What factors influence household energy consumption? They identify household clusters to explore the underlying factors central to understanding electricity consumption behavior. Different clusters carry specific contextual nuances needed for fully understanding consumption behavior. The findings indicate electricity can be distributed according to the needs of six distinct clusters and that utilities can use analytics to identify load profiles for greater energy efficiency.

KEYWORDS

Consumption Pattern, Data Mining, Modeling Energy Consumption, Smart Grid, Smart Meter

INTRODUCTION

Rising electricity consumption has increased fossil fuel production and emissions, with negative environmental impacts (Hinrichs and Kleinbach, 2012). In 2013, residential energy accounted for 29% of the United Kingdom's total energy consumption (DECC, 2013). However, utility providers can use information systems, analytics, "smart grids," and "demand-side management" to accurately forecast electricity consumption and costs, increase productivity while reducing consumption, and enhance their financial bottom line while reducing negative environmental impacts (Corbett, 2013; Nishant et al., 2014; Gholami et al., 2016).

The smart grid is a green IT artifact that can be used to reduce environmental pollution (Corbett, 2010), while demand-side management involves several IT artifacts such as smart meters and meter data management systems to focus on downstream consumption-end activities related to the value chain, with the objective of understanding, influencing, and managing consumer demand (Canever et al., 2008). Demand-side management strategies involve demand response activities such as electricity

DOI: 10.4018/JGIM.2021030109

pricing incentives that can motivate behavior changes (Albadi and El-Saadany, 2008; Gholami et al., 2020) toward more energy efficiency and better load management (Corbett, 2013). Energy efficiency programs focus on reducing the energy use of specific appliances (Energy Information Association, 2011). The programs are typically long term and do not explicitly time demand. Instead, they can motivate efficient consumption patterns and reduce energy consumption by substituting more efficient equipment and providing consumers with specific products suited for their demands (Corbett, 2013; Fritz et al., 2017).

Load management attempts to balance demand with supply in "real time" often involves demand-response programs where consumers may find their electrical services interrupted or changed based on various real-time signals such as price, availability, and grid conditions (Energy Information Association, 2011). Given the application of 15-minute smart meter readings, smart meters provide 35,000 load data points annually. By identifying typical customers, utilities may be better able to forecast energy demands and differentiate among procurement strategies (Fritz et al., 2017). The data gathered through IT is central to the success of any initiative for reducing electricity consumption by providing information that service providers can utilize to design incentives and users can use to modify their consumption.

Studies of energy consumption behavior have had a limited focus. Household energy consumption research is typically one-dimensional (Abrahamse et al., 2005). For example, intervention studies from a psychological perspective (Olander and Thøgerson, 1995) tend to focus predominantly on changing individual-level attitudes. However, equally important are macro-level demographic or societal factors contributing to household energy use and shaping the physical infrastructure that conditions behavioral choices and energy consumption (Abrahamse et al., 2005). Factors such as dwelling characteristics (e.g., floor size and housing type), socioeconomic characteristics (e.g., age of residents) and behavioral characteristics (e.g., appliance usages) are all vital for understanding and forecasting variations of domestic electricity consumption.

The emergence of green IS and energy informatics has led to increased focus on electricity consumption. *Green IS* indicates the use of information systems to promote environmental sustainability (e.g., Elliot, 2011, Jenkin et al., 2011, Melville, 2010, Watson et al., 2012). *Energy Informatics* proposes that energy consumption should be coupled with advanced information systems to improve energy efficiency and reduce emissions (Watson et al., 2010a). Electricity demand forecasting relies extensively on historical data related to factors such as weather patterns, economic conditions, prices, and customer behaviors (Hyndman and Fan, 2009).

Most early demand-side research focused on the outcomes of using smart meters (Chou et al., 2017; Dehdarian 2018; Hielscher and Sovacool, 2019; Kuo et al., 2018; Murray et al., 2018). Although customer attitudes and responses to smart meters determine the ultimate success of demand-side management, the research tells only part of the story (Corbett, 2013). Smart meters also provide utilities with a large amount of new data and information processing capacities that can be used in demand-side management, an information processing activity requiring collection, analysis, and dissemination of information (Corbett, 2013; 2018).

Demand-side management uses advanced metering infrastructure (AMI) systems to measure, collect, transmit, and analyze energy usage. AMI uses smart meters to measure electricity consumption in time intervals, load control switches, and bidirectional communication streams between utilities and consumers (Fridgen et al., 2016). As such, utilities can remotely control demand by emitting control signals that defer electricity consumption in times of higher supply or lower demand, a process called "load shifting" (Fridgen et al., 2016). Drawing on organizational information processing theory and the concept of information waste, Corbett (2013) developed and tested two competing hypotheses using hierarchical regression and data from 543 electricity utilities in the United States.

The model incorporated four levels of metering devices and explained a high portion of the variance in energy efficiency effects of demand-side management programs. The findings indicated that smart meters offer IS-enabled information processing capacities that significantly impact the effectiveness of demand-side management for utilities. The examination included factors within utility control (e.g., metering technology and investments) but not factors related to customer control. In sum, past studies focusing on demand-side management have been predominantly restricted to factors that utilities can manage and manipulate, but have ignored factors specific to residential characteristics.

Studies such as Bale et al. (2015) have also viewed energy consumption in general from a more complicated complexity perspective. When energy systems are viewed as a network, their performance is dependent on both hubs and links in the network (Shang, 2014)[1]. In our studied context, households can be viewed as hubs. As our study is centered around residential energy consumption, we restrict our focus to households or hubs only.

Previous studies focused on electricity consumption often presume that households consume similar amounts of energy and electricity, but we identify household clusters to explore the underlying factors central to understanding general and specific electrical consumption behavior. Different clusters carry specific contextual nuances needed for fully understanding consumption behavior.

Therefore, we address two research questions: *RQ1) Does a data mining approach outperform traditional statistical approaches for modelling residential energy consumption? RQ2) What factors influence household energy consumption?* Our main objective is to clarify the variations and underlying determinants of electricity consumption in the residential sector by building data mining and statistical models. By using data obtained from smart meters (a key green IT asset), we depart from research that uses conventional regression approaches to show how datasets can be leveraged for management studies.

We propose an alternative approach to modeling residential electricity consumption and identifying the underlying determinants. We use a two-step clustering technique to identify household clusters, and then employ regression analysis to tease out the underlying factors of residential electricity consumption. We also examine the effectiveness of the data mining models in investigating the underlying factors of residential electricity consumption. We explore whether segmenting the population into groups and subsequent regression models will improve model performance beyond conventional regression approaches and identify whether household differences regarding dwelling, socioeconomic, and behavioral characteristics cause significantly different electricity consumption.

Our research complements studies highlighting the business, environmental, and social value of using analytics (Constantiou and Kallinikos, 2015; Jui-Sheng et al., 2014; Mithas et al., 2013; Sancho-Asensio et al., 2014; Sharma et al., 2014). More specifically, we make **four** contributions to the literature regarding the role of IT and analytics in addressing societal challenges.

First, unlike past studies that mainly focused on utility-managed factors, we focus on factors specific to consumer-managed residential characteristics. Energy companies will find information systems providing specific feedback on individual households to be extremely valuable for designing targeted motivational policies in energy efficiency campaigns (Loock et al., 2013). Utility providers can better match the demand and supply sides by using insights into customer-specific consumption through smart meters combined with analytics.

Second, researchers have lacked advanced metering technologies and have thus mainly used low-resolution and aggregated energy consumption data. Smart meters now provide disaggregated and high resolution electricity consumption data for a more accurate, refined exploration of the underlying determinants of customers' actual energy consumption.

Third, during the last decade, various areas and industries have used data mining techniques to find that they outperform traditional statistical methods (Swan and Urgusal, 2009). However, we know little about whether data mining techniques can be used to identify the underlying determinants of residential electricity consumption. Fritz et al. (2017) recommended that large utilities would find it economically feasible to use big data solutions, and that small and medium-sized utilities should invest in more cost-effective solutions such as cluster-based systems. By transferring, processing, storing, and analyzing data on a large scale, utility providers could gain valuable information about how to segment load profiles according to consumption patterns among their customer base. Sodenkamp et al. (2015) argued that large industries and energy conservation campaigns would be more effective if they classified customers based on electricity consumption data.

In this study, we use data mining techniques to disaggregate households according to dwelling, socioeconomic, and behavioral characteristics to more clearly identify the underlying determinants of residential electricity consumption and demonstrate the value of data mining for understanding and resolving social issues.

Fourth, previous studies have collected partial data, but our large dataset comprises varied household, socioeconomic, and behavioral characteristics essential for understanding and improving residential electricity consumption. We show that the characteristics interact to determine electricity consumption. For example, the age of residents interacts with their heating usage, which opens possibilities for improving energy efficiency (Abrahamse et al., 2005).

Next, we review studies of IT/IS usage for revealing consumption behavior and literature regarding approaches and research methods used to model household consumption. In the following sections, we present our empirical findings and discuss the implications. In final section, we discuss our conclusions, limitations, and suggestions for future research.

LITERATURE REVIEW

An Overview of The Consumption Behavior Research in IS

IS studies often focus on how IT/IS contributes to consumption of products and services, continued use, continued use intentions, and satisfaction with specific IT/IS artifacts (Bhattacherjee, 2001; Heijden, 2004; Venkatesh et al., 2016). For example, Gelenbe and Caseau (2015) examined IT impacts on energy consumption and emissions. Other studies have focused mostly on the technicalities of specific IT artifacts such as networking architecture (Rawat and Reddy, 2017) and software defined networks (Tuysuz et al., 2017). Our distinct focus on smart meters as an aid for managing electricity consumption departs from conventional IS studies focused on consumption of specific IT artifacts or on how advancements such as online banking or e-government enhance the consumption of specific services. That is, we analyze the influences of the smart meter, a specific IT/IS artifact intended to show users how to better manage their electricity consumption and to help utilities improve energy efficiency.

An Overview of The Approaches Used In Modelling Residential Energy Consumption

Energy economics and energy policy studies have used various approaches for modeling residential energy consumption. Top-down approaches have been useful for forecasting residential energy consumption, but have failed to explain variances, our major objective. We explore bottom-up approaches based on engineering methods to show that they similarly fail to explain underlying determinants. We then review bottom-up statistical and data mining approaches and explain how we address the gaps in bottom-up statistical and data mining research.

Top-Down Models

Swan and Urgursal (2009) explained that models based on top-down approaches forecast the residential sector's energy consumption by estimating total residential energy consumption and variables related to climate and macroeconomic indicators, such as GDP, unemployment rates, and housing growth rates. Therefore, they function at an aggregated level and usually intend to explain how the energy sector relates to the economy. Most top-down models use only a few variables rather than a wide range of factors. Furthermore, "the top-down approach treats the whole residential sector as an energy sink and does not distinguish energy consumption due to individual end-use" (Swan and Urgusal, 2009, p.4). Therefore, models using the top-down approach cannot identify the underlying determinants of energy consumption (Marvuglia and Messineo, 2011).

Bottom-up Models

Bottom-up approaches work at a disaggregated level: energy consumption is calculated for individual or groups of houses and then extrapolated to represent regional or national energy consumption. By considering individual end-users, bottom-up approaches are deemed effective for disclosing how individual determinants impact residential energy consumption (Swan and Urgursal, 2009).

Engineering bottom-up models show how appliances affect energy consumption based on power ratings, appliance use, heat transfer, and thermodynamic principles. However, they fail to incorporate socioeconomic information such as household income and dwelling type (Aydinalp et al., 2001).

In contrast, statistical bottom-up models forecast energy consumption by analyzing energy consumption data from various household characteristics, including occupancy behavior (Ou, 2012) and macroeconomic variables such as income and climate (Swan and Urgusal, 2009).

Statistical Bottom-Up Models

Yohannis et al. (2007) monitored 27 households in Northern Ireland using smart meters on a half-hourly basis for 12 months and found that dwelling type, location, ownership, size, household appliances, number of occupants, income, age, and occupancy patterns all significantly impacted electricity consumption. Pearson correlation statistics showed a clear strong positive linear correlation between floor area and electricity consumption. However, the study failed to provide details such as the average number of hours per day that kitchen, bathroom, and entertainment appliances were used, and the number of loads per day used in washing machines and dishwashers. The study was also limited in size and in averaging electricity consumption for each household over a year. However, climate affects usage of some appliances. For example, air-conditioning is used only in the summer. Overall, *exploratory data analysis* can identify the drivers of residential electricity consumption, but identifying the importance of each factor is much more difficult.

In contrast, *regression analysis* can help identify the importance of predictors. Chen et al. (2012) used multiple linear regression to examine electricity consumption in 1480 residential buildings in Hangzhou, China. Like Yohannis et al. (2007), Chen et al. (2012) collected data on household and socioeconomic characteristics such as the type of dwelling, floor area, age of the main occupant, household income, appliance ownership, and frequency of using appliances, which provided a more insightful examination of underlying determinants.

Similar to Yohannis et al. (2007), Chen et al. (2012) examined residential electricity consumption for more than one season, but selected different participants for each season, carried out two surveys, and collected data via utility bills, which can compromise accuracy by estimating only the amount of energy consumed in each household. In contrast, smart meters can accurately record the amount of energy each household has consumed (Irwin et al., 2012).

O'Dohery et al. (2008) collected a large sample and data range for a more complete examination of electricity consumption, but failed to collect data on the usage or number of appliances and could not identify which appliances consumed the most electricity. Genjo et al. (2004) did not collect data on the frequency or duration of appliance usage, but did incorporate an appliance index to provide some details about household appliances and included household and socioeconomic data such as family income and floor size.

Cramer et al. (1985) incorporated an *appliance index* in a study of 192 California dwellings during the summer season. The study included information about the frequency of appliance use, ownership, location in the dwelling, average efficiencies, and data on household demographic and socioeconomic characteristics. Although the study discovered some underlying determinants of residential electricity consumption, the data were collected through utilities because smart meters had not been invented.

Bedir et al. (2012) tested three regression models. The first was based on duration of appliance use and the presence of occupants (direct determinants); the second was based on the number of appliances and on household and socioeconomic information (indirect determinants); the third combined the variables that emerged as significant predictors in the first two models. However, the models were limited in that the samples were small, and the data collection method was outdated in that electricity providers asked occupants to send their meter readings once a year.

Regression Models To Identify Clearer Insights

To analyze electricity consumption load features, Kavousian et al. (2013) built separate regression models for daily maximum (peak) and minimum (idle) consumption and collected the widest range of household, socioeconomic, and behavioral data such as insulation levels and daily outdoor temperature. From February 28 to October 23, 2010, they used smart meters to gather electricity consumption data at half-hour intervals from 1628 U.S. households. Analyzing data for more than one season can lead to biased findings, but they incorporated a weather indicator variable into their regression models.

They found that location and floor area were among the most important determinants of electricity consumption. In addition, number of refrigerators and entertainment devices most strongly determined daily minimum consumption; number of occupants and high-consumption appliances such as electric water heaters most strongly determined daily maximum consumption. Overall, weather and physical characteristics of buildings had a more significant impact than variables such as occupant behavior.

"Using techniques such as the standard OLS regression are not particularly suitable to target conservation policies towards high energy consumers" (Kaza, 2010, p.2). Instead, quantile regression analysis was used to identify the effects of various factors affecting energy consumption rather than focusing on the total or average electricity consumption. Results showed that housing size and housing type were considerably different at the tails of the energy consumption distribution, while income cases differed by a factor of six.

Regression Models and Multicollinearity

Regression modelling is considered useless if the data contain multicollinearity in which one or more predictors are essentially linear combinations of other predictors (Duntenam, 1989). Among the studies discussed above, only Bedir et al. (2012) and Fuks and Salazar (2006) reported the absence of multicollinearity using the Durbin-Watson test. A few studies have opted to use data mining techniques to reduce the number of predictors and multicollinearity.

Data Mining Segmentation Models

Studies have also used data mining segmentation techniques to identify factors impacting residential energy consumption. For example, Baker and Rylatt (2008) used a clustering technique to identify distinct groups according to behavioral, household, and socioeconomic characteristics. Their primary objective was to determine distinct effects of the two-step clustering technique in small mid-terraced, end-terraced, or larger mid-terraced houses. They found apparent differences between total floor area, occupancy, dwelling age, number of rooms, number of bedrooms, and home office use in each sample, but found little information about whether the clusters consumed electricity differently.

Data Mining Classification Models

In order to estimate building energy performance, the model target variable was conveyed in energy use intensity (EUI), defined as the ratio of annual total energy use to total floor area (the annual total energy use is calculated as the sum of the energy content of all fuel used by the building). Similar to Fuks and Salazar (2006), Yu et al (2010) trained their model using a dependent variable (energy use) in categorical form. However, Yu et al (2010) did initially comprise of a dependent variable in interval form, yet since decision trees work best with the target variable being categorical, the energy use intensity was categorized into a two grade descending scale, i.e. high level and low level, corresponding to low energy performance and high energy performance.

Yu et al (2010) had conducted field surveys for 80 residential buildings located in six different districts in Japan; ten parameters of data (which can be grouped in 4 categories) were collected including climatic conditions (annual average temperature), building characteristics (construction and house type), household characteristics (number of occupants etc.) and household appliance energy sources (kitchen, space heating and hot water supply).

They found outside air temperature was the most important determinant of energy demand, with electric space heating, family size, and house type also all were important determinants of energy demand. Furthermore, the results in Yu et al. (2010) study demonstrated that the use of decision tree method can classify and predict building energy demand levels accurately (93% for training data and 92% for test data), with "HIGH EUI" misclassified as "LOW EUI" three times, whilst "LOW EUI" misclassified as "HIGH EUI" only one time.

Therefore since the results indicated that high EUI is more prone to be misclassified than low EUI, Yu et al (2010) suggested this could be due to the fact that most of the data records are in Low EUI so the decision tree is made more sensitive to this class, therefore an even share between high EUI and Low EUI class in the database would have possibly helped to obtain sufficient accuracy in the two classes.

However, the major limitation of Yu et al (2010) was primarily having in itself only two classes of energy use intensity. Similar to Fuks and Salazar (2006), a large degree in the detail and distinction of results is compromised by possessing a target variable (whether it is electricity consumption or energy use intensity) that is in categorical form, particularly with only two categories as in Yu et al (2010) case. Note that Yu et al (2010) did acknowledge having more than two conceptual levels of energy use intensity can result in a more detailed description, however they felt because of a small sample size, results maybe prone to a high misclassification rate and thus reduce the accuracy of the decision tree.

Neural networks have also been employed to model energy consumption, however, many previous studies have only concentrated on utilizing neural networks via a top-down approach, as Aydinalp et al. (2001, p. 4) states, "*the application of NN has been mainly limited to utility load forecasting*". However Aydinalp et al (2003) applied neural networks in a bottom-up approach. However instead of forecasting electricity consumption as a whole, artificial lighting consumption (ALC) was forecasted using data from the Canadian residential sector. Data was obtained from 988 households and the data included house construction, space heating/cooling and domestic hot-water heating equipment, household appliances, weather and some socio-economic characteristics. Data on energy use was collected from utilities rather than smart meters. The results of this study showed that domestic hot-water heating equipment was the most important predictors of ALC, with a very high prediction rate of 83% occurring.

Method Gaps in The Literature

Previous studies are limited mainly in analyzing limited explanatory variables, such as appliance ownership. However, understanding the relationship between factors, such as between income and appliance load, has considerable potential for improving energy efficiency (Abrahamse et al., 2005). To address the gap, we use a comprehensive dataset covering a wide range of household, socioeconomic, and appliance information for 2035 households in Ireland. Previous studies were also limited in using small samples, using utility bills rather than smart meters to collect energy consumption data, and collecting data from a population of similar household characteristics. To avoid multi-collinearity, reduce the explanatory variables, and better understand energy consumption patterns, we use data mining techniques, clustering technique, and statistical models in a two-stage analysis.

RESEARCH METHOD

Data Sources

Our main data source was the Commission of Energy Regulation (CER), which regulates the electricity and natural gas sectors in Ireland. More than 5000 Irish homes and businesses participated in the CER's Smart Metering Electricity Customer Behavior Trials (CBTs) which assessed the impact of smart meter installation on electricity consumption and introduced the national smart meter rollout.

The CER provided anonymized detailed data underlying the customer behavior trial. They obtained electricity consumption data from the smart meters and obtained data on dwelling, socioeconomic, and behavioral characteristics from a survey of 2035 households. No households surveyed used renewable energy sources.

In late 2007, the CER established the Smart Metering Project Phase 1 in which they conducted smart metering trial runs to assess costs, benefits, and information needed for the full rollout of an optimally designed national smart metering plan. The phase 1 project included technology trials, customer behavior trials, and a cost-benefit analysis for the rollout. The Residential Customer Behavior Trial indicated that time-of-use tariffs and other demand side management stimuli could cause residential trial participants to reduce their overall and peak usage of electricity energy. The Irish Customer Behavior Trial is one of the largest and most statistically robust smart metering behavioral trials conducted internationally to date and thus provides much insight into the impact of smart metering enabled initiatives. A pre-trial survey of participants provided basic allocation information for observing subsequent changes in attitudes, equipment, or electricity usage in the January 2011 post-trial survey.

The main software used to perform data cleaning, data preparation, and modelling was IBM SPSS Modeler, a data-mining and text analytics software for building predictive and prescriptive models and containing modeling options for cluster analysis and rule mining functions.

Data Analysis Steps

We used a two-phase method. First, we used a two-step clustering technique to identify household groups of similar dwelling, socioeconomic, and behavioral characteristics. Next, we used regression analysis for each household group identified in the clustering stage to determine the most important variables and examine how well they explained electricity consumption.

Chiu et al. (2001) developed the two-step cluster analysis. First, pre-clusters are formed to reduce the size of the matrix. The algorithm determines whether each case should be merged with a previously formed pre-cluster or whether the distance between two clusters requires a new pre-cluster. When pre-clustering is complete, all cases in the same pre-cluster are treated as a single entity. The size of the distance matrix depends on the number of pre-clusters rather than the number of cases. In the second step, the standard hierarchical clustering algorithm is employed on the pre-clusters.

The two-step clustering technique has advantages over other clustering techniques. One is the automatic selection of the number of clusters, which solves the problem of not knowing how many clusters to select. Two-step clustering comprises two relative measures of goodness-of-fit: Akaike's information criterion (AIC), which is well-known for overestimating the number of segments, and Bayes information criterion (BIC), which tends to underestimate the number (Berk, 2008). Therefore, both criteria should be used for comparing clustering outcomes.

Two-step clustering has limitations such as excluding respondents with any missing value, thereby reducing sample size. However, in our study, we have a sufficient sample size and thus, our selected approach of two-step clustering does not lead to any empirical issue.

Moreover,two-step clustering was designed specifically to handle both continuous and categorical variables, as in our dataset. The clustering model uses log-likelihood as the distance measure for datasets containing both categorical and continuous variables, as opposed to Euclidean distance which applies to continuous variables only. Also, two-step clustering is better than other clustering techniques such as hierarchical clustering for handling large datasets.

The next step is to build an explanatory model using regression analysis for each cluster identified in the first stage to determine their most significant variables and determine how well they explain the household's total electricity consumption. Once regression coefficients are acquired, the equation can then be used to explain the value of a continuous output (target) as a linear function of one or more independent inputs.

Table 1. Descriptive statistics after SQRT transformation (N=1483)

	Total Electricity Consumption (kw/h)	Approx. Floor Area (square meters)	Age of the House (years)	Approx. Floor Area (square meters) *after SQRT Transformation and Removal of Outliers*
Mean	16.717	44.651	5..629	44.684
Min	3.913	13.631	1.414	28.284
Max	35.833	69.282	9.381	62.048
Range	31.920	55.651	7.967	33.764
Variance	46.904	49.431	2.922	47.671
Stdev	6.849	7.031	1.710	6.832
Skewness	0.287	0.041	-0.232	0.102

Data Pre-Processing

Our 2009 collection of smart meter electricity consumption data at half-hour intervals for 2035 households began on January 9 and ended on February 1. However, some households had end dates before February 1, and data transmission problems caused a few households to have missing intervals of electricity consumption. For 1631 households, January 12 to 26 was a full two-week period with an equal number of week and weekend days. The survey data included 51 variables (Appendix A). We dropped several outlier households that showed high skewness. A square root transformation improved the distribution of the three continuous variables: total electricity consumption, approx. floor area, and age of house, with the skewness values now being 0.393, 0.041, and -0.232 (Table 1).

However, using numerical methods for detecting outliers that are 3 standard deviations from the average and identifying extreme values that are 5 standard deviations from the average revealed that 8 outliers and 0 extreme values were still present in the transformed continuous variable called approx. floor area SQRT. However, no outliers or extreme values occurred in the variables total electricity consumption or age of home. The remaining outliers were discarded, leaving 1483 data points for the analysis (Table 1, last column).

Data Exploration

Appendix A shows a full list of the variables and their ranges. We used graphical analysis and descriptive statistics to determine which variables might be ineffective in the clustering and/or explanatory phase and found that several categorical variables were unrelated to electricity consumption, would not be useful separator variables in the segmentation phase of our methodology, and should be omitted from analysis: Q21) Timer to control when your heating goes on and off? Q22) Timer to control when hot water/immersion comes on or goes off? Q12) Gender of the chief income earner? Q30) How many electric cookers do you have? and Q34) How many immersions do you have?

Research Findings

First, we will explain the results of the clustering phase and the differentiation effectiveness for the clustering model. We will then explain which variables are most important for differentiating each cluster. Then we will explain how each cluster shows significant differences in electricity consumption. Finally, we present regression results.

Two-Step Clustering

We built two clustering models; one based on the AIC goodness-of-fit and the other based on the BIC goodness-of-fit. Both models yielded the same outcomes regarding cluster differentiation, cluster characteristics, and predictor importance. Figure 1 presents the summary of two-step clustering with the Bayes information criterion (BIC) goodness-of-fit. The upper part of the output indicates the quality of our cluster solution. The silhouette measure of cohesion and separation is a measure of the clustering solution's overall goodness-of-fit, essentially based on the average distances between the objects, which can vary between -1 and 1.

Specifically, a silhouette measure of less than 0.20 indicates a poor solution; a measure between 0.20 and 0.50 indicates a fair solution; and values more than 0.50 indicate a good solution. Our measure indicates a satisfactory, almost good, cluster quality. Six clusters of households were identified (Figure 1): the first contains 13.5% of the households; the second contains 15.6%; the third contains 19.2%; the fourth contains 9.1%; the fifth contains 22.8%; and the sixth contains 19.8% (Figure 1 and Table 5 in the appendix).

Regarding the predictor importance of the variables used in the clustering model, the most important were the "duration of appliance use," such as "Q51) How long do you use the gaming consoles?"; "Q50) How long do you use the laptops?"; "Q49) How long do you use the desktop computer?" and "Q48) How long do you use your TVs?" The weakest variables in the clustering model were the "presence of occupants" variables such as "Q18) How many household members under age 15 are typically in the house for about 5 to 6 hours during the day?" and "Q15) How many people over age 15 live in your home?"; and the continuous variable "age of home SQRT."

Comparison of Electricity Consumption Among Clusters

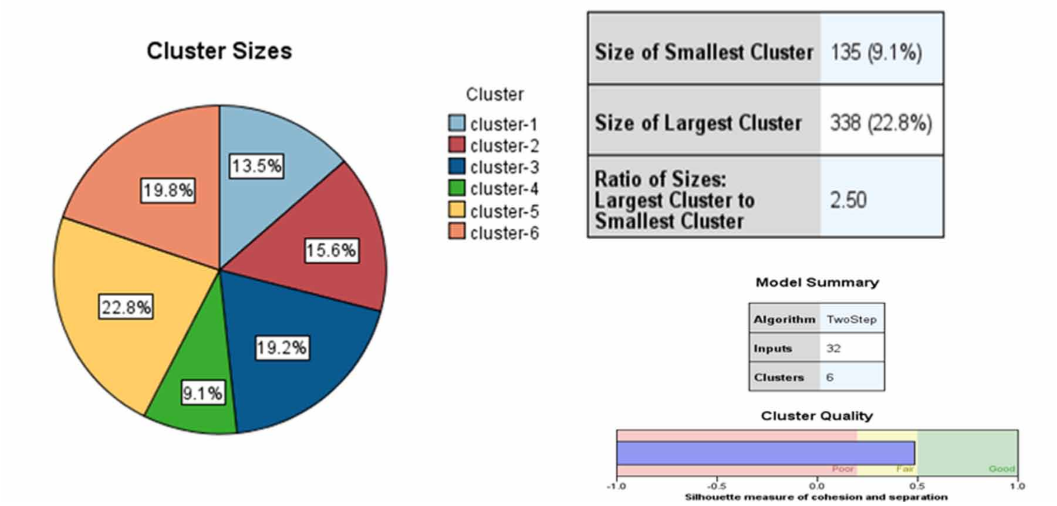
The descriptive statistics indicated that the "total electricity consumption SQRT" for clusters 1 and 4 were negatively skewed, with skewness values of -0.617, and -0.464, while clusters 2, 3, 5, and 6 were positively skewed with skewness values of 0.144, 1.287, 0.646 and 0.062. A Shapiro-Wilko test showed that "total electricity consumption SQRT" was not normally distributed for clusters ($p<0.05$) (Table 2). However, the sample size for each cluster is greater than 50, and the Shapiro-Wilko test flags even minor deviations from normality to be statistically significant. Hence, we used graphical methods such as histograms in addition to the Shapiro-Wilko test to illustrate the results.

The histograms of "total electricity consumption SQRT" confirmed that the distributions were skewed. No outliers or extreme values were found in any clusters. However, since the distribution of "total electricity consumption" in each cluster is not normally distributed, we performed the non-parametric version of the one-way ANOVA test (the Kruskal Wallis), to determine whether statistically significant differences occurred between the distributions of the six clusters.

We performed pairwise comparisons using Dunn's (1964) procedure with a Bonferroni correction or multiple comparisons. Statistical significance was accepted at the $p < .05$ level for the omnibus test and $p < .0083$ level (because of six comparisons, 0.05/6 =0.083) for the multiple comparisons. "Total electricity consumption SQRT" was significantly different between clusters, $\chi^2(5) = 590.109$, p = <.05.

As Figure 2 shows, post-hoc analysis revealed statistically significant differences in "total electricity consumption SQRT" between cluster 1 (median = 24.317) and cluster 2 (median = 18.478) (p=0.00), cluster 1 (median = 24.317) and cluster 3 (median = 12.144) (p=0.00), cluster 1 (median =40) and cluster 5 (median = 49.19) (p=0.36), cluster 1 (median =24.317) and cluster 6 (median = 16.874) (p=0.36), cluster 2 (median = 18.478) and cluster 3 (median 12.144) (p= 0.00), cluster 2 (median = 18.478) and cluster 5 (median 11.098) (p= 0.00), cluster 2 (median = 18.478) and cluster 6 (median 16.874) (p= 0.00), cluster 3 (median = 12.144) and cluster 4 (median 24.038) (p= 0.00), cluster 4 (median = 24.038) and cluster 5 (median 11.098) (p= 0.00), cluster 4 (median = 24.038) and cluster 6 (median = 16.874) (p= 0.00), cluster 5 (median = 11.098) and cluster 6 (median = 16.874) (p= 0.00). However, the analysis showed no statistically significant difference between cluster 1 (median= 24.317) and cluster 4 (median= 24.038) or between cluster 3 (median = 12.144) and cluster 5 (median = 11.098).

Figure 1. Two step clustering summary



Size of Smallest Cluster	135 (9.1%)
Size of Largest Cluster	338 (22.8%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	2.50

Model Summary

Algorithm	TwoStep
Inputs	32
Clusters	6

Table 2. Normality test for total electricity consumption SQRT for each cluster

Cluster	Kolmogorov-Smirnov[a]			Shapiro-Wilko		
	Statistic	df	Sig.	Statistic	df	Sig.
2	0.134	200	0	0.905	200	0
5	0.061	232	0.039	0.971	232	0
3	0.145	285	0	0.867	285	0
4	0.134	135	0	0.936	135	0
6	0.194	338	0	0.926	338	0
1	0.143	293	0	0.899	293	0

Second Stage: Regression Analysis

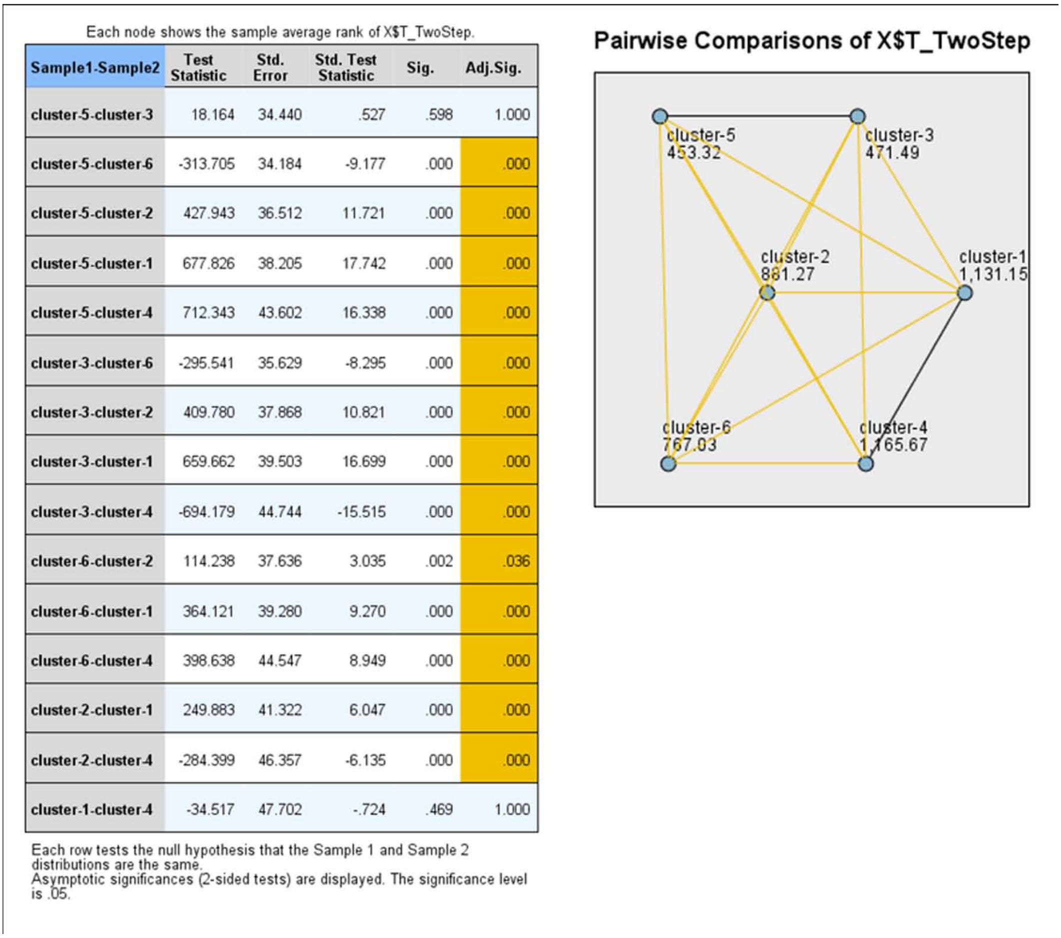
We tested for normality assumption and removed outliers and extreme values before conducting regression analysis. Appendix C presents the summary of assumptions for each regression model. We ran six regression model, as our stage one identified six clusters. Our dependent variable was electricity consumption and we examined how well our explanatory variables explain electricity consumption in each cluster.Table 3 summarizes the statistically significant variables in each regression model, along with the adjusted R^2 values and F-test results for each.

DISCUSSION AND ANALYSIS OF RESULTS

Step 1: Segmentation

Initial segmentation shows how dwelling, socioeconomic, and behavioral characteristics differ among household groups. Cluster 1 represents the ***"lowest socioeconomic level"***. The chief residents are primarily unemployed and have only secondary to intermediate level educations; the families are fairly large; and the accommodations are usually rented small apartments with two bedrooms. Households

Figure 2. Pairwise comparisons



Each node shows the sample average rank of X$T_TwoStep.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
cluster-5-cluster-3	18.164	34.440	.527	.598	1.000
cluster-5-cluster-6	-313.705	34.184	-9.177	.000	.000
cluster-5-cluster-2	427.943	36.512	11.721	.000	.000
cluster-5-cluster-1	677.826	38.205	17.742	.000	.000
cluster-5-cluster-4	712.343	43.602	16.338	.000	.000
cluster-3-cluster-6	-295.541	35.629	-8.295	.000	.000
cluster-3-cluster-2	409.780	37.868	10.821	.000	.000
cluster-3-cluster-1	659.662	39.503	16.699	.000	.000
cluster-3-cluster-4	-694.179	44.744	-15.515	.000	.000
cluster-6-cluster-2	114.238	37.636	3.035	.002	.036
cluster-6-cluster-1	364.121	39.280	9.270	.000	.000
cluster-6-cluster-4	398.638	44.547	8.949	.000	.000
cluster-2-cluster-1	249.883	41.322	6.047	.000	.000
cluster-2-cluster-4	-284.399	46.357	-6.135	.000	.000
cluster-1-cluster-4	-34.517	47.702	-.724	.469	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Pairwise Comparisons of X$T_TwoStep

within this cluster generally use most appliances for a reasonable amount of time. They do not own expensive entertainment appliances such as laptops or desktop computers or kitchen appliances such as washing machines, dishwashers, or tumble dryers. However, perhaps because the chief residents are unemployed, many possess a gaming console and TV and use them for considerable amounts of time, which could explain why they consume the second highest average amount of electricity among all clusters. The findings for cluster 1 contrast with Yohannis et al. (2007) who suggested that the unemployed consume lower levels of electricity.

Cluster 3, the “***lower class singles***,” is similar in that chief residents are also unemployed, mostly middle-aged, with no formal education, and living alone in rented accommodations, with no household residents under 15. We might presume that unemployed residents would consume more electricity because they spend more time indoors, but households in this cluster consume the lowest average amount of electricity. Many of the households do not have washing machines, tumble dryers, dishwashers, gaming consoles, laptops, desktop computers, and internet access.

Regarding dwelling characteristics, households within this cluster do not own energy saving light bulbs, double-glazed windows, insulated external walls, or attic insulation. Their rented apartments are provided by local council housing. The lack of energy saving measures could also indicate the ineffectiveness of the energy-saving measures because they have the lowest average electricity

Table 3. Summary results from regression analysis

	Items that significantly predicted "total electricity consumption"	**Further Details**
Cluster 1	Q20ii) Do you have internet access and/or broadband in your home? yes, internet access Q13ii) What age were you on your last birthday? 26-35 Q13iii) What age were you on your last birthday? 36-45 Q20i) Do you have internet access and/or broadband in your home? yes, internet access and broadband	F (4, 195) = 123.291, p < .000, adj. R^2 = .711 ***Lower Socioeconomic Level Families***
Cluster 2	Q50iv) How long do you use laptops? 3-5 hours per day typically Q48iv) How long do you use your TVs? 3-5 hours per day typically Q14v) Employment status of chief income earner: Unemployed (not actively seeking work) Q39ii) How many washing machine loads do you do per week? 1 Q26ii) How many tumble dryers do you have? 1 Q37ii) How many laptop computers do you have? 1 Q28iii) How many instant electric showers do you have? 2 Q46iv) How long do you use the water pump? 1-2 hours Q45v) How long do you use the electric plugin heater? Over 2 hours Q26iii) How many tumble dryers do you have? 2	F (9, 231) = 33.613, p < .000, R^2 = .585 ***Middle Class Families***
Cluster 3	Q14v) Employment status of chief income earner: Unemployed (not actively seeking work) Q16iii) How many household members over age 15 are typically in the house during the day? 2 Q45iv) How long do you use the electric heater plugin? 1-2 hours Q48iii) How long do you use your TVs? 1-3 hours per day typically Q48iv) How long do you use your TVs? 3-5 hours per day typically Age of home SQRT	F (6, 278) = 175.441, p < .000, adj. R^2 = .790 ***Lower Class Singles***
Cluster 4	Q51iv) How long do you use games consoles? 3-5 hours per day typically Q10ii) Is your attic insulated? If so, when was the insulation fitted? Yes, more than 5 years ago Q9ii) Approx. proportion of double glazed windows: About a quarter Q7iii) How many bedrooms does your home have? 3	F (4, 134) = 107.393, p < .000, adj. R^2 = .761 ***Young Couples***
Cluster 5	Q13ii) What age were you on your last birthday? 26-35 Q43v) How long do you use the electric shower pump? Over 20 mins Q39ii) How many washing machine loads do you do per week? 1 load per week	F (4, 313) = 506.843, p < .000, adj. R^2 = .866 ***Working Class Singles***
Cluster 6	Q1ii) Which best describes your home? semi-detached house Q45iii) How long do you use the electric plugin heater? 30-60 mins Q15iii) How many people over age 15 live in your home? 2 Q43iii) How long do you use the electric shower pump? 5-10 mins Q19iv) Level of education of chief income earner: secondary to leaving cert level	F (6, 292) = 110.520, p < .000, adj. R^2 = .692 ***Retirees***
Entire data	Q7i) How many bedrooms does your home have? 1 Q32i) How many standalone freezers do you have? 0 Q35i) How many TVs do you have? 0 Q41iv) How many dishwasher loads do you do per week? 3 Q15v) How many people over age 15 live in your home? 4 Q26iii) How many tumble dryers do you have? 2 Q15iv) How many people over age 15 live in your home? 3 Q39iii) How many washing machine loads do you do per week? 2 Q39ii) How many washing machine loads do you do per week? 1 Q43ii) How long do you use the electric shower pump? Less than 5 mins Q14v) Employment status of chief income earner: Unemployed (not actively seeking work) Q50v) How long do you use laptops? More than 5 hours per day typically Q15iii) How many people over age 15 years live in your home? 2 Q18i) How many household members under age 15 are typically in the house during the day? 0 Q49iv) How long do you use desktop computer? 3-5 hours per day typically	F (16, 1482) = 278.049, p < .000, adj. R^2 = .749 ***Whole Sample***

consumption. The results support findings reported in prior studies that households with energy saving measures actually consume the most energy (Cramer et al., 1985).

Cluster 5, "***working class singles***," is like cluster 3 in that many of the households include one resident living in rented mostly one-bedroom apartments and lacking energy saving measures such as double-glazed windows or energy saving light bulbs. However, cluster 5 predominately contains employed residents who usually have higher educational levels. They own few kitchen appliances, but they do own entertainment appliances such as TVs, laptops, and gaming consoles, presumably because they can afford them. However, they use most appliances for a fairly short time and thus consume the second lowest average amount of electricity.

Cluster 2, "***middle-class families***," predominately contains employed residents and families living in detached or semi-detached three/four-bedroom houses they own, with mortgages. Most chief residents are between 36 and 55 years-old and have high educational levels. Many residents are under age 15, often young children, typically in the house during the day. Consequently, appliance ownership and usage is relatively high. Most households own all the entertainment appliances (e.g., TVs, laptops, and gaming consoles), and use them for long daily hours. They also own and fairly extensively use kitchen appliances (e.g. electric cookers, washing machines, and tumble dryers). Many of the households have some energy saving light bulbs and double-glazed windows, and many have insulated attics and external walls. Households in this cluster have the highest average approximate floor area.

In Cluster 4, the "***young couples***," most chief residents are 18 to 25, employed, have high levels of education, mid to high paying jobs, and are most likely planning to have a family. Most households have only two people over 15; no households have residents under 15. Approximately half of the households own the house and have a mortgage. The homes have 2/3 bedrooms, and use energy saving measures. Appliance ownership and usage is relatively high, with most households owning all kitchen and entertainment appliances that they use for a long time every day.

Cluster 6, "***retirees***," comprises mostly retirees over 65 who own their home. In most cases, one or two people reside in the home; no households have residents under 15. All households have some energy saving measures. Although most own few entertainment appliances such as laptops or gaming consoles, they own and extensively use most kitchen appliances. Notably, most use electric plugin heaters for over 2 hours, perhaps because older people tend to seek warmth during the winter.

Overall, the findings suggest that clusters differ significantly in dwelling, socioeconomic, and behavioral characteristics, corresponding with our **second research objective** (**Research Question 2**). Cluster 1, which comprises families of lower socioeconomic levels, consume the second highest average amount of electricity because they occupy their homes for longer hours, but they try to conserve their electricity consumption as much as possible when the chief resident is unemployed.

In contrast, cluster 3, which comprises lower-class, mostly unemployed singles, shows the lowest average electricity consumption among all clusters, perhaps because each household typically has only one resident. Cluster 5, working class singles, shows the second lowest average electricity consumption, perhaps because young single people may be engaged in hobbies and activities outside the home. In contrast, cluster 4, comprising young couples, use many kitchen and entertainment appliances for fairly long durations and thus consume the highest average amount of electricity.

Surprisingly, cluster 6, retirees, use kitchen appliances and electric heaters for long durations but show only a mid-level of electricity consumption, perhaps because they rarely use entertainment appliances. We also tested for statistically significant differences between the average electricity consumption for each cluster and found no statistically significant differences between clusters 1 and 4, and clusters 3 and 5.

Overall, segmentation analysis provided insight into group characteristics that influence electricity consumption, corresponding with our **first research objective** (**Research Question 1**). The clustering highlighted that users are not homogeneous, and therefore each cluster requires separate examination. This suggests that a data mining approach indeed offers more insights compared to traditional statistical approaches for modelling residential energy consumption

Step 2: Regression Analysis

Our second research question was focused on understanding *what factors influence household energy consumption?* Partially confirming achievement regarding our **second research objective**, regression results show that cluster variables, such as appliance usage, significantly predict electricity consumption, but do not explain why certain variables are significant in each regression model. Beyond understanding cluster characteristics and the effects on electricity consumption, we must understand the underlying determinants of electricity consumption. However, the apparent randomness of significant variables makes it difficult to succeed in our explanatory models. Nonetheless, similar to Bedir et al. (2012) and Yau et al. (2005), age of the household and the approximate floor area were among the most important predictors for each cluster.

The first research objective was to examine the proposed model for its effectiveness in predicting residential electricity consumption(**Research Question 1**). The explanatory models achieved overall good results. The adjusted R2 was high for several clusters. The explanatory power for various models varied from 0.585 to 0.866, aligned with previous research that found fairly similar explanatory R^2 rates from 50% to 70% (Bedir et al., 2012; Cramer et al., 1985; Genjo et al., 2004). The explanatory power for a sole regression model without clustering was 0.76. The model with the same variables has different explanatory power for different clusters. Thus a sole regression model would not reveal variations between user groups. Consequently, the findings support the need for clustering the users before conducting regression analysis (**Research Question 1**). Different clusters vary socioeconomically and in appliance consumption, which limits any generic policy prescription or research implication emanating from pooled regression analysis.

Research Implications

In reviewing previous studies, we identified a major limitation in that studies were restricted to partial sets of explanatory variables, such as appliance ownership, an issue identified in past studies such as Shang (2017). Nevertheless, we must understand the relationship between various factors, such as between income and appliance loads, if we are to improve energy efficiency (Abrahamse et al., 2005). Therefore, we used a comprehensive dataset covering a wide range of household, socioeconomic, and appliance information, acquired because the smart meter, an effective, key green IT artifact, provides new and more comprehensive datasets for addressing environmental problems (Corbett 2010; 2018; Watson et al., 2010). Data acquisition precedes data analytics and is essential for understanding energy consumption and encouraging sustainability.

Previous research was also limited in using small samples, using utility bills to collect energy consumption data, and applying “concealed” techniques such as factor analysis that fail to provide detailed, clear results. Instead, we used smart meters to collect data from a population of similar household characteristics. We adopted a data analytics methodology combining data mining and statistical techniques to provide the details needed to interpret results. Our study therefore suggests that using the prolific data mining approach in statistical analysis is a useful complement to statistical analysis. That is, data mining techniques allowed us to disaggregate households with different dwelling, socioeconomic, and behavioral characteristics, and therefore more clearly identify the underlying determinants of residential electricity consumption.

Companies could create business, environmental, and social value by using IT and big data analytics (Constantiou and Kallinikos, 2015; Fritz et al., 2017; Jui-Sheng et al., 2014; Loock et al., 2013; Mithas et al., 2013; Sancho-Asensio et al., 2014; Sharma et al., 2014; Sodenkamp et al., 2015). Our study suggests some interesting nuances. For instance, some clusters were strong consumers of electric appliances, but not computing devices. Thus, we revealed that IT studies should consider a variety of electric appliances.

Specifically, smart devices such as appliances that utilize sensors should be considered beyond traditional computing devices if energy efficiency measures are to be successful. Counterintuitively, consuming traditional rather than modern electric devices may be more environmentally friendly. Thus the behavioral perspective is more important than the technical perspective and may negate potential benefits of new technologies.

We therefore call for approaching green IS from a behavioral rather than technical perspective, based on socioeconomic characteristics. Studies regarding successful IS are often restricted to aspects such as satisfaction, intentions to continue using, or actual use. Smart meters are unique in helping users and utility service providers reduce energy consumption and carbon footprints. We show the complex underlying policy and behavior factors that may determine the success of IT artifacts. Thus, we need to conceptualize different measures of success.

We also add to the general literature showing that firms use business analytics to curate relevant data, perform basic analysis, and integrate analytics into current processes and that IT artifacts can reveal potential, ability, and context (Tim et al., 2019). Similarly, we integrate analytics into conventional statistics. We also demonstrate that IT artifacts combined with data mining potential provides better consideration of context, as evident in our nuanced results.

Managerial and Policy Implications

Smart meters enable electricity utilities to capture billions of data points regarding demand according to conditions, customer segments, and time of use (Corbett, 2013). Rather than using a one-size fits-all approach, utilities can use the information to design and implement a more segmented and effective demand-side management portfolio, which requires the two-way communication that is particularly enhanced through the smart grid (Corbett, 2013; Farhangi, 2010; Strueker and Dinther, 2012; Valocchi et al., 2007). By deploying smart meters, utilities can develop a robust communication channel for communicating with customers and allowing utilities to send price or supply signals as part of demand-response management (Strueker and Dinther, 2012).

Utilities can formulate specific policies for each cluster group identified. For instance, they could concentrate on reducing the amount of electricity consumed by entertainment appliances for clusters 1 (lower class families) and 2 (middle class families)[2] since both show high use of entertainment appliances such as TVs, gaming consoles, and computers. Utilities could formulate policies to reduce the electricity consumed by comfort appliances such as showers and electric plugin heaters for clusters 4 (young couples) and 6 (retirees). Overall, our segmentation/regression-based methodology allowed us to explore the underlying determinants of electricity consumption in the residential sector.

Innovative utility providers are finding competitive advantages in leveraging big data analytics (Hazen et al., 2016), in precisely segmenting their customer base according to characteristics that determine the use of products and services (McGuire et al., 2012). Big data analytics is no longer a new idea requiring validation. Instead, it is an increasingly necessary strategic reality for competing in the data-grounded economy and rapidly changing competitive marketplace (Hazen et al., 2016).

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

In this study, we used segmentation/regression to identify the underlying determinants of electricity consumption in the residential sector. The segmentation phase primarily identified some underlying determinants. The regression models provided acceptable explanatory rates for each cluster indicating that bottom-up models can predict long-term electricity consumption in the residential sector.

However, we were unable to show that double-glazed windows and energy saving light bulbs were particularly effective for saving energy, because households using energy saving measures still consumed high amounts of electricity (rebound effect). Rebound effect is the percentage of energy savings from efficiency that are offset by increased use. Efficiency makes an energy-consuming technology less expensive to use, so people use it more often so a 10% improvement in efficiency might provide only a 9% reduction in energy use (Thomas and Azevedo, 2013; Latiner 2000; Schipper, and Grubb, 2000).

Future research should identify the significance of the different underlying determinants in each cluster, perhaps by making a few modifications to the proposed methodology. First, collecting appliance usage data in a continuous format would be more effective. Incorporating continuous variables in predictive models such as regression can clarify the importance of each variable. For example, how much Kwh is consumed by an extra 10 minutes of TV use?

Finally, as an alternative to regression and neural network models, the second phase of the segmentation/regression approach could use decision trees to provide easily interpreted results. While decision trees have advantages has such as it is intuitive and easy to explain, it is considered less appropriate for continuous variables (Dhiraj, 2019). Beyond dwelling, socioeconomic, and behavioral characteristics, variables such as weather and economic data can also improve the findings. In summary, we used data from smart meters, a key green IT artifact, coupled with data mining and statistical analysis, to understand electricity consumption behavior among various groups. By showing that smart meter usage outperforms sole regression approaches, our research has extensive research and policy implications.

ACKNOWLEDGMENT

We thank the Irish Social Science Data Archive (ISSDA) for providing the data used for this project (www.ucd.ie/issda). We also thank Mr. Navdeep Shergill, a postgraduate student in Operations & Information Management Group at Aston Business School, UK, for helping with this project.

REFERENCES

Abrahamse, W., Steg, L., Vlek, C., & Rothengatter, T. (2005). A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, *25*(3), 273–291. doi:10.1016/j.jenvp.2005.08.002

Aydinalp, M., Ugursal, V., & Fung, A. (2004). Modelling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks. *Applied Energy*, *79*(2), 159–178. doi:10.1016/j.apenergy.2003.12.006

Baker, K., & Rylatt, M. (2008). Improving the prediction of UK domestic energy-demand using annual consumption data. *Applied Energy*, *85*(6), 400–431. doi:10.1016/j.apenergy.2007.09.004

Bale, C. S., Varga, L., & Foxon, T. J. (2015). Energy and complexity: New ways forward. *Applied Energy*, *138*, 150–159. doi:10.1016/j.apenergy.2014.10.057

Bedir, M., Hasselaara, E., & Itard, L. (2013). Determinants of electricity consumption in Dutch dwellings. *Energy and Building*, *58*(11-12), 194–207. doi:10.1016/j.enbuild.2012.10.016

Berk, R. (2008). *Statistical Learning from a Regression Perspective*. Springer Science.

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *Management Information Systems Quarterly*, *25*(3), 351–370. doi:10.2307/3250921

Chen, J., Wang, X., & Steemers, K. (2012). A statistical analysis of a residential energy consumption survey studying Hangzhou, China. *Energy and Building*, *66*(3), 193–202.

Chou, J. S., Telaga, A. S., Chong, K., & Gibson, G. E. Jr. (2017). Early-warning application for real-time detection of energy consumption anomalies in buildings. *Journal of Cleaner Production*, *149*, 711–722. doi:10.1016/j.jclepro.2017.02.028

Constantiou, I. D., & Kallinikos, J. (2014). New games, new rules: Big data and the changing context of strategy. *Journal of Information Technology*, *30*(1), 44–57. doi:10.1057/jit.2014.17

Corbett, J. (2013). Using information systems to improve energy efficiency: Do smart meters make a difference? *Information Systems Frontiers*, *15*(5), 747–760. doi:10.1007/s10796-013-9414-0

Corbett, J., Wardle, K., & Chen, C. (2018). Toward a sustainable modern electricity grid: The effects of smart metering and program investments on demand-side management performance in the us electricity sector 2009-2012. *IEEE Transactions on Engineering Management*, *65*(2), 252–263. doi:10.1109/TEM.2017.2785315

Cramer, J., Miller, N., Craig, P., & Hackett, B. (1985). social and engineering determinants and their equity implications in residential electricity use. *Energy*, *10*(12), 1283–1291. doi:10.1016/0360-5442(85)90139-2

DECC. (2013). Overall energy consumption in the uk since 1970. In DECC energy consumption in the UK. London: National Statistics

Dehdarian, A. (2018). Scenario-based system dynamics modeling for the cost recovery of new energy technology deployment: The case of smart metering roll-out. *Journal of Cleaner Production*, *178*, 791–803. doi:10.1016/j.jclepro.2017.12.253

Dhiraj, K. (2019). *Top 5 advantages and disadvantages of Decision Tree Algorithm*. https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a

Elliot, S. (2011). Transdisciplinary perspectives on environmental sustainability: A resource base and framework for IT-enabled business transformation. *Management Information Systems Quarterly*, *35*(1), 197–236. doi:10.2307/23043495

Fridgen, G., Häfner, L., König, C., & Sachs, T. (2016). Providing utility to utilities: The value of information systems enabled Flexibility in electricity consumption. *Journal of the Association for Information Systems*, *17*(8), 538–563. doi:10.17705/1jais.00434

Fritz, M., Albrecht, S., Ziekow, H., & Strüker, J. (2017), Benchmarking big data technologies for energy procurement efficiency. *American Conference on Information Systems Proceedings*.

Fuks, M., & Salazar, E. (2008). Applying models for ordinal logistic regression to the analysis of household electricity consumption classes in Rio de Janeiro, Brazil. *Energy Economics*, *30*(2), 1672–1692. doi:10.1016/j.eneco.2007.09.006

Gelenbe, E., & Caseau, Y. (2015). The impact of information technology on energy consumption and carbon emissions. *Ubiquity*, *2015*(June), 1–15. doi:10.1145/2755977

Gholami, R., Emrouznejad, A., Alnsour, Y., Kartal, H. B., & Veselova, J. (2020). The Impact of Smart Meter Installation on Attitude Change towards Energy Consumption Behavior among Northern Ireland Households. *Journal of Global Information Management*, *28*(4), 21–37. doi:10.4018/JGIM.2020100102

Gholami, R., Watson, R. T., Hasan, H., Molla, A., & Anderson, N. B. (2016). Climate Change and Green IS Solutions: How can we do more? *Journal of Association for Information Systems*. *Special Issue on Information Systems Solutions for Environmental Sustainability*, *17*(8), 308–313.

Hazen, B. T., Skipper, J. B., Ezell, J. D., & Boone, C. A. (2016). Big Data and predictive analytics for supply chain sustainability: A theory-driven research agenda. *Computers & Industrial Engineering*, *101*, 592–598. doi:10.1016/j.cie.2016.06.030

Heijden, V. D. H. (2004). User acceptance of hedonic information systems. *Management Information Systems Quarterly*, *28*(4), 695–704. doi:10.2307/25148660

Hielscher, S., & Sovacool, B. K. (2018). Contested smart and low-carbon energy futures: Media discourses of smart meters in the United Kingdom. *Journal of Cleaner Production*, *195*, 978–990. doi:10.1016/j.jclepro.2018.05.227

Hinrichs, R., & Kleinbach, M. (2012). *Energy: Its Use and the Environment* (5th ed.). Brooks/Cole.

Jenkin, T. A., Webster, J., & McShane, L. (2011). An agenda for 'green' information technology and systems research. *Information and Organization*, *21*(1), 17–40. doi:10.1016/j.infoandorg.2010.09.003

Jui-Sheng, C., Yu-Chien, H., & Liang-Tse, L. (2014). Smart meter monitoring and data mining techniques for predicting refrigeration system performance. *Expert Systems with Applications*, *41*(5), 2144–2156. doi:10.1016/j.eswa.2013.09.013

Kavousian, A., Ram, R., & Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, *55*, 184–194. doi:10.1016/j.energy.2013.03.086

Kaza, N. (2010). Understanding the spectrum of residential energy consumption: A quantile regression approach. *Energy Policy*, *38*(2), 6574–6585. doi:10.1016/j.enpol.2010.06.028

Kuo, T. C., Tseng, M. L., Lin, C. H., Wang, R. W., & Lee, C. H. (2018). Identifying sustainable behavior of energy consumers as a driver of design solutions: The missing link in eco-design. *Journal of Cleaner Production*, *192*, 486–495. doi:10.1016/j.jclepro.2018.04.250

Laitner, J. A. (2000). Energy efficiency: Rebounding to a sound analytical perspective. *Energy Policy*, *28*, 6–7, 471–475.

Loock, C.-M., Staake, T., & Thiesse, F. (2013). Motivating Energy-Efficient Behavior with Green IS: An Investigation of Goal Setting and the Role of Defaults. *Management Information Systems Quarterly*, *37*(4), 1313–1332. doi:10.25300/MISQ/2013/37.4.15

Marvuglia, A., & Messineo, A. (2009). A. Using Recurrent Artificial Neural Networks to Forecast Household Electricity Consumption. *Energy Procedia*, *14*(45), 430–462.

McGuire, T., Manyika, J., & Chui, M. (2012). *Why big data is the new competitive advantage. Ivey Business Journal*. July/August.

Melville, N. (2010). Information systems innovation for environmental sustainability. *Management Information Systems Quarterly*, *34*(1), 1–21. doi:10.2307/20721412

Mithas, S., Lee, M. R., Earley, S., Murugesan, S., & Djavanshir, R. (2013). Leveraging big data and business analytics. *IEEE IT Professional*, *15*(6), 18–20. doi:10.1109/MITP.2013.95

Murray, D. M., Stankovic, L., Stankovic, V., & Espinoza-Orias, N. D. (2018). Appliance electrical consumption modelling at scale using smart meter data. *Journal of Cleaner Production*, *187*, 237–249. doi:10.1016/j.jclepro.2018.03.163

Nishant, R., Teo, T. S. H., & Goh, M. (2014). Energy Efficiency Benefits: Is Technophilic Optimism Justified? *IEEE Transactions on Engineering Management*, *61*(3), 476–487. doi:10.1109/TEM.2014.2314703

Rawat, D. B., & Reddy, S. R. (2017). Software defined networking architecture, security and energy efficiency: A survey. *Environment*, *3*(5), 6–20.

Sancho-Asensio, A., Navarro, J., Arrieta-Salinas, I., Armendáriz-Íñigo, J.-E., Jiménez-Ruano, V., Zaballos, A., & Golobardes, E. (2014). Improving data partition schemes in Smart Grids via clustering data Streams. *Expert Systems with Applications*, *41*(13), 5832–5842. doi:10.1016/j.eswa.2014.03.035

Schipper, L., & Grubb, M. (2000). On the rebound? Feedback between energy intensities and energy uses in IEA countries. *Energy Policy*, *28*(6-7), 6–7, 367–388. doi:10.1016/S0301-4215(00)00018-5

Shang, Y. (2014). Vulnerability of networks: Fractional percolation on random graphs. *Physical Review. E*, *89*(1), 012813. doi:10.1103/PhysRevE.89.012813 PMID:24580287

Shang, Y. (2017). Subgraph robustness of complex networks under attacks. *IEEE Transactions on Systems, Man, and Cybernetics. Systems*, *49*(4), 821–832. doi:10.1109/TSMC.2017.2733545

Sharma, R., Mithas, S., & Kankanhalli, A. (2014). Transforming decision-making processes: A research agenda for understanding the impact of business analytics on organizations. *European Journal of Information Systems*, *23*(4), 433–441. doi:10.1057/ejis.2014.17

Sodenkamp, M., Kozlovskiy, I., & Staake, T. (2015). Gaining is business value through big data analytics: a case study of the energy sector. *Proceedings of the International Conference on Information Systems*.

Swan, L., & Ugursal, V. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable & Sustainable Energy Reviews*, *1*(13), 1819–1835. doi:10.1016/j.rser.2008.09.033

Thomas, B. A., & Azevedo, I. L. (2013). Estimating direct and indirect rebound effects for U.S. households with input–output analysis Part 1: Theoretical framework. *Ecological Economics*, *86*, 199–210. doi:10.1016/j.ecolecon.2012.12.003

Tim, Y., Hallikainen, P., Pan, S. L., & Tamm, T. (2019). Actualizing business analytics for organizational transformation: A case study of Rovio Entertainment. *European Journal of Operational Research*. Advance online publication. doi:10.1016/j.ejor.2018.11.074

Tso, G., & Yau, K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, *32*(9), 32. doi:10.1016/j.energy.2006.11.010

Tuysuz, M. F., Ankarali, Z. K., & Gözüpek, D. (2017). A survey on energy efficiency in software defined networks. *Computer Networks*, *113*, 188–204. doi:10.1016/j.comnet.2016.12.012

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2016). Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the Association for Information Systems*, *17*(5), 328–276. doi:10.17705/1jais.00428

Watson, R. T., Boudreau, M.-C., & Chen, A. (2010a). Information systems and environmentally sustainable development: Energy informatics and new directions for the IS community. *Management Information Systems Quarterly*, *34*(1), 1–16. doi:10.2307/20721413

Watson, R. T., Corbett, J., Boudreau, M. C., & Webster, J. (2012). An information strategy for environmental sustainability. *Communications of the ACM*, *55*(7), 28–30. doi:10.1145/2209249.2209261

Yohannis, Y., Mondol, J., Wright, A., & Norton, B. (2006). Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use. *Energy and Building*, *40*(6), 1053–1059. doi:10.1016/j.enbuild.2007.09.001

Yu, Z., Fariborz, H., Fungb, B., & Yoshinoc, H. (2010). A decision tree method for building energy demand modelling. *Energy and Buildings*, *1*(1).

ENDNOTES

[1] We would like to thank an anonymous reviewer for guiding us to this aspect.

[2] https://www.sciencenewsforstudents.org/article/couch-potatoes-tend-be-tv-energy-hogs

APPENDIX A

Table 4.

Item	Type	Range	Strength	Relationship with *"Total Electricity Consumption"*
Electricity Consumption				
Total Electricity Consumption SQRT	Continuous	-	-	-
Dwelling Characteristics				
Q1) Which best describes your home?	Nominal	1-5	Medium	Apartments generally consume lower amounts of electricity. Terraced and detached houses consume higher amounts.
Q2) Do you own or rent your home?	Nominal	1-4	Medium	Residents who rent homes from local authorities consume the lowest electricity. Home owners with or without mortgages consume the most.
Age of the house SQRT	Continuous	-	Medium	Strong positive linear relationship, Pearson correlation of 0.793.
Floor area SQRT	Continuous	-	Low	Weak relationship, Pearson correlation of 0.294.
Q7) How many bedrooms are in your home?	Nominal	1-5	Medium	One-bedroom households generally consume low amounts of electricity. Five or more bedroom households consume the highest amounts.
Q8) Approx. proportion of energy saving light bulbs?	Nominal	1-5	Medium	Households that use energy saving light bulbs surprisingly consume the most electricity.
Q9) Approx. proportion of double-glazed windows?	Nominal	1-5	Medium	Households with all proportions of double-glazed windows surprisingly consume the most electricity. Households that have only have a quarter proportion or no double-glazed windows consume the least.
Q10) Is your attic insulated? If so, when was it fitted?	Nominal	1-3	Medium	Households with attic insulation consume generally high amounts of electricity, while households lacking attic insulation consume both low and high amounts.
Q11) Are the external walls insulated?	Nominal	1-2	Medium	Households with insulated external walls consume medium to high amounts of electricity, while some households lacking insulated external walls consume low amounts.
Socioeconomic Information				
Q12) Gender of the chief income earner?	Nominal	1-2	None	Gender was not associated with electricity consumption and was omitted from the analysis.
Q13) What age were on your last birthday?	Nominal	1-6	Low	Age was not associated with electricity consumption, but the normalized histogram shows that chief income earners have evenly spread ages. Therefore, the variable could be an important separator of households in the clustering phase of our methodology.
Q14) Employment status of chief income earner?	Nominal	1-7	Medium	Chief income earners who are unemployed but actively seeking work consume low amounts of electricity.
Q15) How many people over age 15 live in your house?	Nominal	1-5	Strong	The more residents living in households, the greater the electricity consumption. Number of residents can be discerned from the normalized histogram.
Q16) How many people over age 15 are typically in the house during the day?	Nominal	1-5	Medium	The greater the number of residents over age 15 living in the house during the day, the greater the electricity consumption.
Q17) How many people under age 15 live in your house?	Nominal	1-6	Medium	Many households in the dataset have no residents under age 15. Therefore, this categorical variable might not be an important separator of households in the clustering phase. However, the normalized histogram shows that the greater the number of residents under age 15, the greater the electricity consumption.
Q18) How many people under age 15 are typically in the house during the day?	Nominal	1-6	Medium	The conclusions are similar to the variable Q17 above.
Q19) Level of education of chief income earner?	Nominal	1-5	Low	Chief income earners who lack formal education tend to consume less electricity.
Appliance Information				
Q20) Do you have internet access and/or broadband in your home?	Nominal	1-3	Medium	Most households without internet access consume low levels of electricity. Households with internet access and broadband consume the highest levels.
Q21) Timer to control when your heating goes on and off?	Nominal	1-2	None	Electricity consumption varies equally for households with and without timers to control heating and hot water/immersion. The variable is not associated with electricity consumption and was omitted from analysis.
Q22) Timer to control when hot water/immersion comes on or goes off?	Nominal	1-2	None	Conclusions are similar for Q21 above, so this variable was omitted from analysis.
Q23) Do you use immersion when your heating is not switched on?	Nominal	1-2	Medium	Households that use immersion when the heating is off tend to consume higher levels of electricity than those who do not use the immersion.

continued on following page

Table 4.Continued

Item	Type	Range	Strength	Relationship with *"Total Electricity Consumption"*
Q24) Does your hot water tank have a lagging jacket?	Nominal	1-2	Low	Households that have lagging jackets on their hot water tanks consume the lowest levels of electricity; those who do not have lagging jackets consume the highest levels.
Q25) How many washing machines do you have?	Nominal	1-3	Medium	Households that have no washing machine consume lower amounts of electricity than those that have one.
Q26) How many tumble dryers do you have?	Nominal	1-3	Medium	Conclusions are similar to those for the variable Q25 above.
Q27) How many dishwashers do you have?	Nominal	1-3	Medium	Conclusions are similar to Q25 and Q26 above.
Q28) How many instant electric showers do you have?	Nominal	1-3	Low	Many households lacked instant electric showers, so the variable might be ineffective in the clustering phase. The small proportion of households with electric showers consumed large amounts of electricity, particularly with more than two.
Q29) How many electric showers pumped from the hot tank do you have?	Nominal	1-3	Low	Conclusions are similar to Q28, and the variable might be very effective in the clustering phase. However, households without electric showers pumped from the hot tank consume high amounts of electricity.
Q30) How many electric cookers do you have?	Nominal	1	None	Every household has one electric cooker, so it is not associated with electricity consumption.
Q31) How many electric plugin Heaters do you have?	Nominal	1-3	Medium	Households without electric plugin heaters consume the lowest levels; households with one or two consume the highest levels.
Q32) How many standalone freezers do you have?	Nominal	1-3	Low	Most households own one standalone freezer. The small minority of households owning two consume the most electricity. Therefore, freezers have a small association with electricity consumption.
Q33) How many water pumps do you have?	Nominal	1-3	Low	Many households have only one water pump, so the variable is likely to be ineffective in the clustering phase. However, the few households without water pumps consume the lowest levels of electricity; the few households with two consume high levels.
Q34) How many immersions do you have?	Nominal	1-2	None	Most households have only one immersion, so the variable will be ineffective in the clustering phase. The extremely small minority of households owning two immersions consume the most electricity.
Q35) How many TVs do you have?	Nominal	1-4	Low	Most households have only one TV and consume the lowest amounts of electricity along with those who own no TVs. Those who own two or three consume the highest amounts.
Q36) How many desktop computers do you have?	Nominal	1-3	Low	Most households have no desktop computers, so the variable is most likely ineffective in the clustering phase. However, those who own one desktop computer consume high levels of electricity, while the small minority owning two consumes some of the highest levels.
Q37) How many laptop computers do you have?	Nominal	1-4	Low	Conclusions are similar to Q36 above.
Q38) How many gaming consoles do you have?	Nominal	1-4	Low	Conclusions are similar to Q36 and Q37 above.
Q39) How many washing machine loads do you do per week?	Nominal	1-5	Medium	Households that report no washing machine loads achieve the lowest levels of electricity consumption; those who do three or more washing machine loads per week use the highest levels.
Q40) How many tumble dryer loads do you do per week?	Nominal	1-5	Medium	Conclusions are similar to Q39 above.
Q41) How many dishwasher loads do you do per week?	Nominal	1-5	Medium	Conclusions are similar to Q39 and Q40 above.
Q42) How long do you use the instant electric shower?	Nominal	1-5	Low	Many households do not own an instant electric shower, so this variable will most likely be ineffective in the clustering phase. However, the small minority of households who use the instant electric shower for over 20 minutes consume high amounts of electricity.
Q43) How long do you use the electric shower pumped from the hot tank?	Nominal	1-5	Low	Households using their electric shower pumped from the hot tank for 20 minutes or longer actually consumed lower levels electricity than those who did not the appliance, indicating that an instant electric shower uses more electricity than an electric shower pumped from a hot tank.
Q44) How long do you use the electric cooker?	Nominal	1-4	Medium	Households using electric cookers for long durations have high electricity consumption, while those using cookers for shorter periods use less electricity.
Q45) How long do you use the electric plugin heater?	Nominal	1-4	Medium	Conclusions are similar to Q44 above.
Q46) How long do you use the water pump?	Nominal	1-5	Medium	Conclusions are similar to Q44 and Q45 above.
Q47) Do you use the standalone freezer for all or part of the year?	Nominal	1-2	Medium	Conclusions are similar to Q44, Q45, and Q46 above.

continued on following page

Table 4.Continued

Item	Type	Range	Strength	Relationship with *"Total Electricity Consumption"*
Q48) How long do you use your TVs?	Nominal	1-5	Low	Households using TVs for 3 to 5 hours or more consume high levels of electricity. Those who use TVs for less than 1 hour or 1-3 hours a day consume lower levels. The duration of use has such wide ranges that the variable can be effective in the clustering phase.
Q49) How long do you use your desktop computer?	Nominal	1-5	Low	Households using desktop computers for more than 5 hours consume the highest levels of electricity.
Q50) How long do you use your laptop computer?	Nominal	1-5	Low	Conclusions are similar to Q49 above.
Q51) How long do you use your gaming console?	Nominal	1-5	Low	Households using gaming consoles for less than an hour a day consume lower amounts of electricity; households using gaming consoles for more than five hours per day consume some of the highest amounts of electricity.

APPENDIX 2

Table 5. Characteristics of each cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Q1) Which best describes your home?	94%: terraced house	64.22%: detached house 31.47%: semi-detached house	100%: apartment	100%: semi-detached house	100%: apartment	49.49%: terraced house, 30.38% bungalow
Q2) Do you own or rent your home?	100%: rent from a private landlord	84.91%: own with mortgage	100%: rent from a local authority	52.59%: own with mortgage, 47.41%: rent from a local authority	78.11%: rent from a private landlord	85.67%: own, mortgage free
Q7) How many bedrooms are in your home?	68%: 2 bedrooms 36%: 3 bedrooms	40.95%: 3 bedrooms, 55.17%: 4 bedrooms	89.82%: 1 bedroom	60%: 2 bedrooms, 40%: 3 bedrooms	71.01%:1 bedroom	53.58%: 2 bedrooms, 43%: 3 bedrooms
Q8) Approx. proportion of energy saving light bulbs?	91%: 0	35.34%: about three-quarter proportion 51.72%: all	100%: 0	68.15%: 0 31.85%: about half	100%: 0	46.42%: about three quarter proportion, 51.54%: all
Q9) Approx. proportion of double-glazed windows?	94%: all	83.62%: all	100%: 0	67.41%: all, 31.11%: about a quarter	100%: about a quarter	91.13%: all
Q10) Is your attic insulated? If so, when was it insulated?	76.50%: yes within last 5 years	70.69%: yes within last 5 years	100%: no	74.81%: yes within last 5 years	100%: 0	59.73% yes, more than 5 years ago, 30.38% no
Q11) Are the external walls of your home insulated?	86%: yes	91.38% yes	100% no	92.59% yes	100%: yes	67.92%: yes, 32.08%: no
Q13) What age were you on your last birthday?	44%: 36-45, 31%: 46-55, 25.50%: 26-35	45.69% 36-45 29.74% 46-55	65.96% 36-45, 32.28% 46-55	77.78% 18-25	86.39%: 18-25	77.13%: 65+
Q14) Employment status of chief income earner?	100%: unemployed (not actively seeking work)	72.84%: employed	58.60%: unemployed (actively seeking work), 41.40%: unemployed (not actively seeking work)	88.89%: employed	97.63%: employed	82.25%: retired, 17.75%:carer
Q15) How many people over age 15 live in your home?	68%: 3	60.34%: 2 28.45%: 3	85.61%: 1	82.96%: 2	94.67%: 1	42.66%: 1 37.88%: 2
Q16) How many residents over age 15 are typically in the house during the day?	67%: 3	60.78%: 1 35.78%: 2	83.86%: 2	96.30%: 0	100%: 0	67.92%: 1
Q17) How many people under age 15 live in your home?	23%: 0 23%: 1 18.50%: 3 56%: 4	54.74%: 3 39.22%: 2	99.65%: 0	100%: 0	100%: 0	94.20%: 0
Q18) How many people under age 15 are typically in the house during the day?	23%: 0 21%: 3 54%: 4	65.95%: 2	100%: 0	100%: 0	100%: 0	95.96% 0
Q19i) Level of education of chief income earner?	87.50%: secondary to intermediate junior cert level	60.34%: third level, 28.88%: secondary to leaving cert level	100%: No formal education	63.70%: third level, 32.59%: secondary to leaving cert level	45.56%: secondary to intermediate junior cert level, 33.14%: secondary to intermediate junior cert level, 21.30%: third level	64.85%: secondary to intermediate junior cert level
Q20) Do you have internet access and/or broadband in your home?	42.50%: no, 31.50% yes, internet access 26%: yes, internet access and broadband	95.26%: yes, internet access and broadband	100%: no	100%: yes, internet access and broadband	85.80%: yes, internet access and broadband	80.89%: no
Q23) Do you use your immersion when your heating is off?	91%: Yes	64.22%: Yes	100%: No	88.15%: No	100%: No	81.57%: Yes
Q24) Does your hot water tank have a lagging jacket?	72%: No	87.93%: Yes	100%: No	100%: No	78.11%: No	77.82%: Yes
Q25) How many washing machines do you have?	68.50%: 0 31.50%: 1	98.28%: 1	100%: 0	100%: 1	82.84%: 0	88.74%: 1

continued on following page

Table 5.Continued

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Q26) How many tumble dryers do you have?	89.50%: 0	65.95%: 1 25.43%:0	100%: 0	100%: 1	94.08%: 0	88.05%: 1
Q27) How many dishwashers do you have?	98.50%: 0	87.93%: 1	100%: 0	57.04%: 1 42.96%: 0	100%: 0	61.77%: 1 38.23%: 0
Q28) How many instant electric showers do you have?	98.50%: 0	59.91%: 1 20.69%: 2	100%: 0	82.22%: 1	100%: 0	85.32% 1
Q29) How many electric showers pumped from a hot tank do you have?	100%: 1	80.60%: 0	100%: 1	100%: 0	100%: 1	85.32%: 1
Q31) How many electric plugin heaters do you have?	67.50%: 0 32.50%: 1	93.97%: 1	100%: 0	100%: 1	100%: 0	95.56%: 2
Q32) How many standalone freezers do you have?	100%: 1	68.53%: 1 29.74%: 2	100%: 1	100%: 0	100%: 1	95.56%: 1
Q33) How many water pumps do you have?	100% 1	79.31% 2	100%: 1	100%: 1	59.17%: 1, 40.83%: 0	95.56%: 1
Q35) How many TVs do you have?	85% 1	69.40%: 2 25%: 3	95.44%: 2	66.67%: 1 33.33%: 2	100%: 1	86.35%: 1
Q37) How many laptop computers do you have?	82.50% 0	84.05% 2	100%: 0	91.42% 1	91.42%: 1	95.56%: 0
Q38) How many gaming consoles do you have?	82.50% 0	58.19%: 2 30.17%: 1	99.65%: 0	81.48%: 2	51.18%: 1 48.82%: 2	100%: 0
Q39) How many washing machine loads do you do per week?	68.50%: 0 22%: 2	61.21% More than 3	100%: 0	100%: 1	82.84%: 0	89.76%: More than 3
Q40) How many tumble dryer loads do you do per week?	89.50%: 0	60.34%: 4 33.62%: 3	100%: 0	100%: 1	94.08%: 0	63.82%: 3 29.35%: 2
Q41) How many dishwasher loads do you do per week?	98.50%: 0	67.67%: More than 3 27.59%: 3	100%: 0	66.67%: More than 3 33.33%: 3	100%: 0	48.46%: 3 38.23% 0
Q42) How long do you use the instant electric shower?	98.50%: 0	58.19%: 5-10 mins 29.31%: 0	100%: 0	100%: 0	100%: 0	92.49% 0
Q43) How long do you use the electric shower pump?	59.50%: 5-10 mins, 40.50%: 10-20 mins	70.69%: 0	100%: 5-10 mins	100%: 0	87.87%: Over 20 mins	79.86%: Less than 5 mins
Q44) How long do you use the electric cooker?	89% 30-60 mins	75% Over 2 hours	100%: 30-60 mins	44.67%: Less than 30 mins 28.89%: 30-60 mins, 24.44%: 1- hours	100% Less than 30 mins	79.18% 30-60 mins
Q45) How long do you use the electric plugin heater?	69.50%: 30-60 mins 30.50%: 1-2 hours	62.07%: 1-2 hours 37.93%: Over 2 hours	69.82%: 30-60 mins 30.18%: 1-2 hours	74.07%: 30-60 mins	100%: Less than 30 mins	86.01%: Over 2 hours
Q46) How long do you use the water pump?	93.00%: 30-60 mins	51.72%: Over 2 hours, 30.60% 1-2 hours	100%: 30-60 mins	62.96%: Less than 30 mins, 37.04% 30-60 mins	59.17%: Less than 30 mins, 40.83% None	80.89%: 1-2 hours
Q47) Do you use the standalone freezer for all or part of the year?	83%: part of the year	93.97%: all of the year	100%: part of the year	100%: all of the year	90.24%: all of the year	68.60%: part of the year
Q48i) How long do you use your TVs?	100%: more than 5 hours	51.72%: 1-3 hours, 26.72%: less than 1 hour	95.44%: more than 5 hours	58.52%: 1-3 hours, 24.44%: 3-5 hours, 17.04%: less than 1 hour	100%: less than 1 hour	84.30%: 3-5 hours per day
Q49) How long do you use your desktop computer?	94%: do not use	86.21%: 3-5 hours	100%: do not use	69.63%: less than 1 hour 22.22%: 1-3 hours	100%: less than 1 hour	93.52%: do not use
Q50) How long do you use laptops?	82.50%: do not use	100%: do not use	72.59%: 1-3 hours	72.59%: 1-3 hours	71.30%: 1-3 hours	95.56%: do not use
Q51) How long do you use gaming consoles?	59%: more than 5 hours	82.76%: 3-5 hours	99.65%: do not use	59.26%: 1-3 hours 40.74%: 3-5 hours	51.18%: less than 1 hour, 48.82%: 1-3 hours	100%: do not use
Approx. floor area SQRT (average)	43.374	53.106	40.545	49.191	43.278	42.484
Age of home SQRT (average)	7.338	5.338	4.868	7.385	4.42	6.005
Total electricity consumption (average)	22.962	18.707	11.962	23.667	12.116	17.469

APPENDIX 3

Table 6. Summary of assumptions for each regression model

	Linearity	Independence of observations	Homoscedasticity	Multi-collinearity	Normality
Cluster 1	The residuals formed a somewhat horizontal band.	Durbin-Watson statistic of 1.781 is quite close to 2, so we can accept an independence of observations.	The residuals were fairly equally spread over the predicted values of the dependent variable.	No correlations greater than 0.7, VIF measures for all dependent variables were less than 10.	The mean and standard deviations of the standardized residuals were 0.00 and 0.990, respectively. The residuals appear approximately normal from the histogram. The P-P plot shows that the points are close enough to indicate normality although they are not perfectly aligned along the diagonal and have a slight peak down the diagonal line.
Cluster 2	The residuals formed a somewhat horizontal band.	Durbin-Watson statistic of 1.177 is quite close to 2, so we can accept an independence of observations.	The residuals were mostly equally spread over the predicted values of the dependent variable.	No correlations greater than 0.7, VIF measures for all dependent variables were less than 10.	The mean and standard deviation of the standardized residuals were 0.00 and 0.978, respectively. The residuals appear approximately normal from the histogram. The P-P plot shows that the points are close enough to indicate normality although they are not perfectly aligned along the diagonal and have a slight peak down the diagonal line.
Cluster 3	The residuals formed a somewhat horizontal band.	Durbin-Watson statistic of 1.320 is quite close to 2, so we can accept an independence of observations.	The residuals were fairly equally spread over the predicted values of the dependent variable.	No correlations greater than 0.7, VIF measures for all dependent variables were less than 10.	The mean and standard deviation of the standardized residuals after omitting 6 outliers were 0.00 and 0.989, respectively. After we deleted the outliers, the residuals appear approximately normal from the histogram. The P-P plot shows that the points are close enough to indicate normality although they are not perfectly aligned along the diagonal and have a slight peak down the diagonal line.
Cluster 4	The residuals formed a somewhat horizontal band.	Durbin-Watson statistic of 1.258 is quite close to 2, so we can accept an independence of observations.	The residuals were mostly equally spread over the predicted values of the dependent variable.	No correlations greater than 0.7, VIF measures for all dependent variables were less than 10.	The mean and standard deviation of the standardized residuals were 0.00 and 0.978, respectively. The residuals appear approximately normal from the histogram. The P-P plot shows that the points are close enough to indicate normality although they are not perfectly aligned along the diagonal and have a slight peak down the diagonal line.
Cluster 5	The residuals formed a somewhat horizontal band.	Durbin-Watson statistic of 1.563 is quite close to 2, so we can accept an independence of observations.	The residuals were mostly equally spread over the predicted values of the dependent variable.	No correlations greater than 0.7, VIF measures for all dependent variables were less than 10.	The mean and standard deviation of the standardized residuals, after omitting 24 outliers, were 0.00 and 0.994, respectively. Furthermore, after outliers were deleted, the residuals appear to be approximately normal from the histogram. The P-P plot shows that the points close enough to indicate normality although they are not perfectly aligned along the diagonal and have a slight peak down the diagonal line.
Cluster 6	The residuals formed a somewhat horizontal band.	Durbin-Watson statistic of 1.234 is quite close to 2, so we can accept an independence of observations.	The residuals were mostly equally spread over the predicted values of the dependent variable.	No correlations greater than 0.7, VIF measures for all dependent variables were less than 10.	The mean and standard deviation of the standardized residuals were 0.00 and 0.990, respectively. The residuals appear approximately normal from the histogram. The P-P Plot shows that the points are close enough to indicate normality although they are not perfectly aligned along the diagonal and have a slight peak down the diagonal line.
Entire data	To a small extent, the residuals formed a band, but quite a few residuals were further away, obscuring the horizontal band.	Durbin-Watson statistic of 1.187 is quite close to 2, so we can accept an independence of observations.	The residuals were unequally spread over the predicted values of the dependent variable, causing a negative linear relationship.	No correlations greater than 0.7, VIF measures for all dependent variables were less than 10.	The mean and standard deviation of the standardized residuals were 0.00 and 0.995, respectively. The residuals appear approximately normal from the histogram. The P-P plot shows that the points are close enough to indicate normality although they not perfectly aligned along the diagonal and have a slight peak down the diagonal line.

Rohit Nishant's current research interest includes Green IS/IT, sustainable operations, sustainability, IT to address grand social challenges, business and social value of technology, and e-government. His research has been accepted for publication or published in reputed journals such as MIS Quarterly, Journal of the Association for Information Systems, The Journal of Strategic Information Systems, IEEE Transactions on Engineering Management, MIS Quarterly Executive, Journal of Environmental Management, and Journal of Business Research. Currently, he is working on developing my stream of research on green IS/IT, sustainable operations, and sustainability.

Ali Emrouznejad is a Professor and Chair in Business Analytics at Aston Business School, UK. His areas of research interest include performance measurement and management, efficiency and productivity analysis as well as data mining and big data. He holds an MSc in applied mathematics and received his PhD in operational research and systems from Warwick Business School, UK. Dr Emrouznejad is editor / associate / guest editor / member of editorial boards of several scientific journals including Annals of Operations Reserach, European Journal of Operational Reserach and Socio-Economic Planning Sciences. . He has published over 150 articles. Dr Emrouznejad is editor / author of several books including (1) "Applied Operational Research with SAS" (CRC Taylor & Francis), (2) "Big Data Optimization" (Springer), (3) "Performance Measurement with Fuzzy Data Envelopment Analysis" (Springer), (4) "Managing Service Productivity" (Springer), (5) "Fuzzy Analytics Hierarchy Process" (CRC Taylor & Francis), and (6) "Handbook of Research on Strategic Performance Management and Measurement".