Energy Economics 87 (2020) 104697

Contents lists available at ScienceDirect

Energy Economics

journal homepage: www.elsevier.com/locate/eneco

Understanding segmentation in rural electricity markets: Evidence from India[☆]

Daniel Robert Thomas[a], Shalu Agrawal[b], S.P. Harish[c], Aseem Mahajan[d], Johannes Urpelainen[e],[*]

[a] Columbia University, USA
[b] Council on Energy, Environment and Water, India
[c] College of William and Mary, USA
[d] Harvard University, USA
[e] Johns Hopkins SAIS, USA

ARTICLE INFO

Article history:
Received 28 April 2019
Received in revised form 15 January 2020
Accepted 30 January 2020
Available online 7 February 2020

JEL classification:
Q40
Q41
C38
O13
Q47
L94

Keywords:
Electricity
Energy
Demand
India
Clustering
Segmentation
Rural development

ABSTRACT

How can demand for electricity be estimated without fine-grained usage data? Employing an original and large dataset, we develop a novel method for determining drivers of demand without electricity meter data. We first segment Indian consumers by their willingness to pay for electricity service, their level of usage, and their satisfaction with lighting, and then use cluster membership as a dependent variable in order to determine which household-level factors predict electricity usage. Our approach employs machine-learning and more traditional regression techniques to determine the optimal number of segments, generate the segments, and determine the predictors of segment membership. The dataset consists of more than 10,000 households in more than 200 villages in the states of Bihar, Odisha, Rajasthan, and Uttar Pradesh. We find that the rural Indian electricity market can be segmented into three clusters based on households' willingness to pay, satisfaction with lighting, and appliance wattage. The clusters consist of potential customers, low-demand customers, and high-use customers. We then determine the predictors of membership in these clusters. We show that different types of consumers can be identified along easily observable measures. Moreover, we show that there are clear groups of consumers that vary along their satisfaction, willingness to pay, and existing appliance usage.

1. Introduction

What factors affect consumers' electricity demand in rural India? To create robust electricity markets in rural areas, we first need to clarify why some households value electricity more than others. In energy-poor contexts, a concrete step towards understanding drivers of demand and uptake of electricity is an analysis of associations between household and village-level characteristics, and electricity consumption. With a nuanced understanding of the rural electricity consumer, policymakers can identify and tailor policy interventions that allow rural electricity demand to grow over time. Here, we employ a market segmentation design of Indian consumers from over 200 villages in four states, clustering them by their willingness to pay for electricity service, their level of usage, and their satisfaction with lighting. We then use cluster membership as a dependent variable in order to determine which household-level factors predict electricity usage.

To clarify the intuition of what our approach achieves, imagine an organization trying to design an intervention to increase electricity demand by improving quality of service. This organization would ostensibly be interested in consumers who currently have low demand for electricity (which we measure through willingness to pay) and low satisfaction with service quality. A simple linear model would only be able to capture one of these factors as a dependent

[☆] We thank Ryan Kennedy and Joseph Sutherland for excellent comments. We gratefully acknowledge Smart Power India Foundation and Rockefeller Foundation for funding and partnership.

[*] Corresponding author at: Rome Building, 4th Floor, 1619 Massachusetts Avenue, NW, Washington, DC 20036, USA.

E-mail addresses: daniel.thomas@columbia.edu (D. Robert Thomas), JohannesU@jhu.edu (J. Urpelainen).

https://doi.org/10.1016/j.eneco.2020.104697

variable. However, by creating customer segments, we can create multidimensional outcomes that, in this example, can reflect both low willingness to pay and low satisfaction with service. Covariates predicting membership in this cluster can then be used to target specific types of customers.

Market segmentation is an empirical exercise intended to separate a market of customers or potential customers into heterogeneous groups which can be specifically targeted (Assael and Roscoe, 1976). Proponents of segmentation analysis argue that the structure of modern markets requires that customers' needs and the heterogeneity of their needs are recognized in the production and marketing of goods (Wedel and Kamakura, 2012). Market segmentation occurs in several stages; as variables relevant to the marketing of a certain product must be selected, measured, sometimes converted into factors, and used to create segments (Foedermayr and Diamantopoulos, 2008). Variables used in segmentation are generally differentiated along two dimensions: whether they are macro-level (economic, geographic, cultural) or micro-level (individual-level information), and whether they are directly observable, or inferred. Moreover, the type of segments created can differ, as they can be typologies of consumer behavior, or profiles of consumer segments based on descriptive data (Assael and Roscoe, 1976). Once identified, segments can be a powerful tool for marketing.

For policymakers, segmentation of rural electricity markets is also useful. Without understanding the current and potential demand among rural consumers, effective planning for electricity distribution is all but impossible. Some consumers have high latent demand for electricity, and they could constitute the backbone of a robust electricity market with growing demand. Others haver lower levels of demand, meaning that even minimal amounts of energy would suffice. Given the near-monotonic growth in energy consumption over time in contemporary India, understanding the determinants of demand is becoming increasingly important.[1]

Despite the widespread use of market segmentation, its use in electricity markets and within India has been limited. Indeed, Dutta and Mitra (2015, p. 21) note that "academic research on electricity market segmentation in India and many other economies is rare." These authors maintain that engaging in this research can ultimately lead to better revenue management and the development of suitable pricing schemes for customers. Singh (2007) echoes this call for greater engagement in segmentation research, especially in rural settings, due to variation in demand based on climatic conditions, occupations, literacy and other factors.

To answer this call, we use a multi-pronged segmentation approach in order to both generate and describe segments in the rural Indian electricity market, and to determine which factors predict membership in different segments. Our approach employs both machine-learning and regression techniques to determine the optimal number of segments, generate the segments, and analyze the predictors of segment membership. Major advantages of data-driven clustering is that the analyst need not specify in advance what the customer groups are and skewed data distributions with outliers do not compromise the integrity of the exercise. Our clustering approach avoids the need to hypothesize in advance how to categorize the different dimensions and how many households to place into each. The algorithm produces the number and characteristics of different segments based on the nature of the data itself, in a systematic but inductive approach.

We employ an original data-set for this exercise, consisting of 10,249 households in 204 villages in the states of Bihar, Odisha, Rajasthan and Uttar Pradesh. The villages were chosen sequentially, beginning with a sample of villages with either (i) a mini-grid electricity source or (ii) a private distribution franchise in the state of Odisha. To that sample, similar villages from the same district were added for comparative purposes. Overall, the villages were relatively large to sustain a market with some rural enterprises, yet suffering from low levels and quality of electricity access. While this sampling does not produce a representative sample of rural India, it gives ample variation in electricity sources and quality of service in areas of interest to rural energy planners and enterprises. The data was collected in collaboration with Smart Power India (SPI) in 2018. The questionnaire collected detailed information on demand for and use of electricity at the household level. In the dataset, all but 2 villages are connected to the electrical grid, but not all households within these villages have grid connections.

We find that the rural Indian electricity market can be segmented into three clusters along willingness to pay, satisfaction with lighting, and appliance wattage. Qualitatively, the three clusters represent different levels of potential and observed electricity usage. This differentiation between clusters allows identification of those who might be likely to demand greater electricity in the future, due to potential for increases in satisfaction and current usage. We find a large cluster of customers in what we call the 'potential customer' cluster. Individuals in this cluster use low levels of electricity currently, and have low satisfaction with quality of lighting. This indicates that an improvement in service to these customers could increase their demand. We also identify a large cluster of low-demand customers. In this cluster, current usage is low, but satisfaction with lighting is high. Thus, we argue that these customers will remain at their current usage levels regardless of improvements in service. Most customers fall into these two clusters, with the low-demand cluster composing over half of the dataset. Our third cluster consists of customers with both high satisfaction and high usage, and we label it the high-use cluster. This cluster also likely will exhibit low growth in demand in the future. This cluster is also the smallest in our dataset.

Segmentation variables were chosen to capture different aspects of customers' projected demand, despite not having access to data from metering technology. However, our strategy uncovers important relationships between these variables: for example, we find that the cluster in which customers exhibit the highest willingness to pay and appliance wattage is not the same as the cluster in which customers exhibit the highest satisfaction with lighting. Analysis without segmentation would not capture this relationship in regression results.

Using OLS regression, we identify several strong predictors of cluster membership. An increase in the age of respondents predicts their membership in the low-demand cluster, while reducing the likelihood that they are in the potential customer cluster. Increases in the education level of respondents increase the likelihood they are in the high-use cluster, while reducing the likelihood they are in the potential customer cluster. Increases in the size of households predicts membership in the high-use cluster, while it decreases the likelihood of being in the low-demand cluster. Finally, being in a scheduled caste or tribe reduces the likelihood of being in the low-demand and high-use clusters, while increasing the likelihood of being in the potential customer cluster.

Psychographic variables, used to describe customers on psychological attributes, are also significantly related to cluster membership. Having leadership traits and being risk averse decrease the likelihood of being in the potential customer cluster, while increasing the likelihood of being in the high-use cluster. Having leadership traits also predicts membership in the low-demand cluster. Finally, preferring cheap goods increases the likelihood of being in the low-demand cluster while reducing the likelihood of being in the high-use cluster.

Finally, we estimate models with endogenous variables included. These results should be interpreted cautiously. However, we do find

[1] For data on patterns of energy consumption in the last few years in India, see Gaur et al. (2016).

that as the number of hours of grid electricity increases in a village, its inhabitants are more likely to be in the low-demand and high-use clusters. Increases in household expenses decrease the likelihood of being in both the potential customer and low-demand clusters while increasing the likelihood of being in the high-use cluster. Finally, having a mini-grid or grid connection decreases the likelihood of being in the potential customer cluster and increases that of being in the low-demand cluster. Having a grid connection increases the likelihood of being in the high-use cluster.

These results offer a first-step towards characterizing the rural Indian electricity market, and offer direction for electricity providers and NGOs attempting to market their electricity products. We show that different types of consumers can be identified along easily observable measures. Moreover, we show that there are clear groups of consumers that vary along their satisfaction, willingness to pay, and existing appliance usage. The implications for policymakers and non-governmental organizations of this research are clear: generating easy to measure data on observable characteristics of potential consumers can lead to a powerful ability to target consumers who are most willing to demand electricity. Moreover, this same strategy can allow organizations to avoid the increased provision of energy to those who do not seek it, leading to a more efficient allocation of electricity overall.

2. Electricity segmentation

Our approach contributes to the understanding of demand for electricity in rural India, but also speaks to a larger literature. In particular, we take a step forward in the use of segmentation for estimating demand in electricity markets, and energy markets in general.

There has been limited engagement in market segmentation research for energy markets in general. Analyzing willingness to pay for green electricity in China, Zhang and Wu (2012) use a multinomial logit model to identify relevant variables for segmentation and then create market segments. They find that demographic variables predict willingness to pay. Similarly, Hyland et al. (2013) estimate the gross margin earned from the supply of electricity to households in Ireland, and find that they can be predicted by economic and demographic variables. In the context of the energy market in India, Srinivasan (2005) adopts a qualitative approach towards segmenting the photovoltaic market along several dimensions. However, few other papers have engaged in this research, despite its usefulness for energy policy. Indeed, in the context of energy, Rajabi et al. (2017) argue that clustering can be useful for tariff design, load forecasting, demand response and customer classification. In this study, we employ clustering for customer classification.

In India itself, market segmentation research is in its infancy. Kiran Mor (2014) segments the Indian market using a random sample of 400 consumers, using a psychographic approach. Similarly, Narang (2011) uses psychographic data on Indian youths to identify clusters in apparel store selection. However, this is one of the first studies to segment the rural Indian market, and to our knowledge, the first to segment the electricity market in India.

Our approach speaks to a larger literature on access and usage of electricity in rural India, and the use of energy sources in general, specifically work which determined factors predicting the adoption of new energy technologies. Recent work on energy adoption in rural India shows that household expenditures and savings, along with entrepreneurial attitudes, are key determinants of innovation and the adoption of new technology (Aklin et al., 2018). This confirms earlier findings that wealth drives adoption in at least some types of energy markets (Smith and Urpelainen, 2014). In a review of 32 studies concerning the adoption of improved fuels and cookstoves, Lewis and Pattanayak (2012) showed that income and education predicted adoption, whereas the effect of variables such as household size, composition and gender were unclear. We test similar household-level factors in our analysis, providing clarity on their role.

Previous work on the determinants of rural electrification in India has largely been located at the state or regional level, attributing poor electrification to structural factors or poor governance (Palit and Chaurey, 2011). In their study of determinants of rural electrification in Bihar, Oda and Tsujita (2011) employ village-level data, and show that location is the most important determinant of electrification. Our study uses household-level data in order to determine the factors that influence the adoption of electricity within villages. This level of analysis can explain variation in access to electricity, as “in spite of having moderate to high village electrification rate, the household connections in rural … India continue to be low” (Palit and Chaurey, 2011, p. 272).

Moreover, past analysis at the household-level in rural India has been limited. Bhattacharyya (2006) used household-level data to conclude that electricity consumption increases with higher levels of income and in urban areas in India, but presented only descriptive analysis. Kemmler (2007) improved on Bhattacharyya 's (2006) design by employing a binary choice model, and found that household electrification depended on household characteristics, the level of community electrification, and the quality of electricity supply. However, this analysis was limited by the data available, in that the dependent variable was limited to whether or not a household was electrified. We build upon these results by employing a richer dataset specific to energy usage by which we can better measure variation in the consumption of electricity.

In general, our approach differs from past studies by combining a segmentation approach to defining demand profiles, and modeling the factors that predict demand. Our paper generates descriptives segments of the Indian rural electricity market, similar to past attempts at segmenting energy markets. However, it takes a step forward by modeling segment membership, giving rich quantitative analysis in addition to qualitative characterization of the segments.

3. Research design

Our approach consists of several steps. To begin, we choose variables with which to segment the data. Variables were chosen in order to create clusters of individuals at different consumption levels. Using these variables, we then determine the optimal number of clusters in the data using an unsupervised machine learning technique. After determining the number of clusters, we use the k-means clustering algorithm to create clusters of our data according to three variables. Once these clusters are created, we employ individual-level cluster membership as a dependent variable and test which household and village-level factors predict membership in the high-use cluster using OLS regression. This approach is useful in two ways in that it allows us to describe the market for rural electricity in India, and determine which variables predict households' consumption of electricity.

3.1. Data

The data for the study were collected in collaboration with SPI, an organization established by the Rockefeller Foundation. The survey was constructed to study technology adoption, power consumption and customer attitudes towards mini-grid and grid electricity in rural India. The dataset consists of 10,249 households in 204 villages in four states: Bihar, Odisha, Rajasthan and Uttar Pradesh. To our knowledge, this is one of the largest and most fine-grained surveys regarding the rural market for electricity in India.

SPI previously implemented mini-grid and distribution franchise interventions in this area. We sampled from both intervention and non-intervention villages. First, we randomly sampled 50 mini-grid and distribution franchise villages. Then, we randomly sampled 50 villages similar to the mini-grid intervention villages, and 50 similar to distribution franchise intervention villages.

The sample for the survey randomly selected villages that were similar to SPI intervention villages along two covariates: total population and distance to nearest town. Villages that were within one standard deviation of the mean of these covariates for the intervention villages were included in the sample frame, reducing expected sampling variability.

The questionnaire was designed to gather detailed information on electricity usage and demand at the household level. To this end, we collected fine-grained data on methods of electrification, willingness to pay for electricity, and appliance ownership and usage, among other variables. The detail of the data allows for an understanding of the factors driving electrification at a more micro-level than past studies. Moreover, the large sample size and sample frame create strong external validity for the study, allowing for generalizability.

3.2. Segmentation variables

We use three variables to segment the data. Variables were chosen in order to create clusters of individuals at different consumption levels. The first is willingness to pay.[2] In our survey, we ask respondents about their willingness to pay for four different combinations of electricity service. For the purposes of this analysis, we use the highest level of service, in which respondents were asked to name how much they would pay for “uninterrupted electricity throughout the day, which allows you to use all of your electric appliances.”[3] In the Appendix S1, we show the correlation between the four willingness to pay packages measured in the questionnaire. The four packages are highly correlated so we restrict our analysis to employing only one.

The second variable used in segmentation is total appliance wattage. In the questionnaire, we asked respondents about 20 specific types of appliances: we asked if they owned the appliance, if so, how many of the appliance they owned, the wattage of the appliance, and the brand of the appliance. This allowed us to construct a detailed measure of the total wattage used by appliances within households. This variable is constructed by multiplying the number of devices owned by households by their wattage. Further details on the assumptions made to construct this variable are available in Appendix Section S6. Wattage is intended to capture demand by households. We did not survey households without electricity on their appliance ownership, and thus treat their wattage as 0. To account for missingness in the wattage data, we imputed data as follows: first, we limited the extreme values of appliance wattage by creating upper and lower bounds through secondary research and consultations with appliance shop owners. For customers who were unsure of their appliance wattage, we used the mean value within the bounds identified. Furthermore, for certain appliances which generated a substantial amount of missingness during piloting, typical values from secondary research were used. These appliances included washing machines, music or radio systems, TV sets, mobile phones and laptops.

The third variable used in segmentation is satisfaction with lighting. Satisfaction is measured on a 5-point scale, with 1 indicating that the respondent is very unsatisfied with their lighting arrangement, and 5 indicating that they are very satisfied. We anticipate that satisfaction with lighting may capture a critical component of future demand of electricity. Namely, if attitudes towards outcomes of electricity supply, such as artificial lighting, factor into a customer’s demand, then those with similar levels of satisfaction may be likely to demand similar levels of service in the future. We further expect a strong correlation between satisfaction and willingness to pay, given the existing experimental research (Homburg et al., 2005).

We show the correlation between the three segmentation variables in Fig. 1. Appliance wattage and willingness to pay are correlated, whereas satisfaction with lighting is not correlated with the other variables.

These three variables allow us ample variation with which to determine different clusters of consumers in the rural Indian electricity market. Moreover, the variables proxy for different types of engagement with electrification, in terms of value, usage, and satisfaction. Thus, clustering across these variables allows us to profile specific types of possible consumers. The following section describes our approach to determining predictors of cluster membership.

3.3. Predictor variables

We employ a multitude of variables across several models to explain cluster membership. We first begin with village and individual-level socioeconomic variables. Namely, we test the effects of age, education level, religion (as a binary variable capturing whether a respondent is Hindu), caste (as a binary variable measuring whether a respondent is part of a backward caste or tribe), household size, and the number of hours of grid electricity at the village level. Testing associations between these individual level variables is important for several reasons. first, these attributes are most easily measured and available to those seeking to provide electricity. Thus, understanding their association with demand can improve marketing and expansion strategies. Second, these variables capture important variation in households’ remaining lifetime earning potential, social mobility and other factors that may be correlated with demand. Our predictor variables are in line with past literature on demand for electricity (Jones et al., 2015).

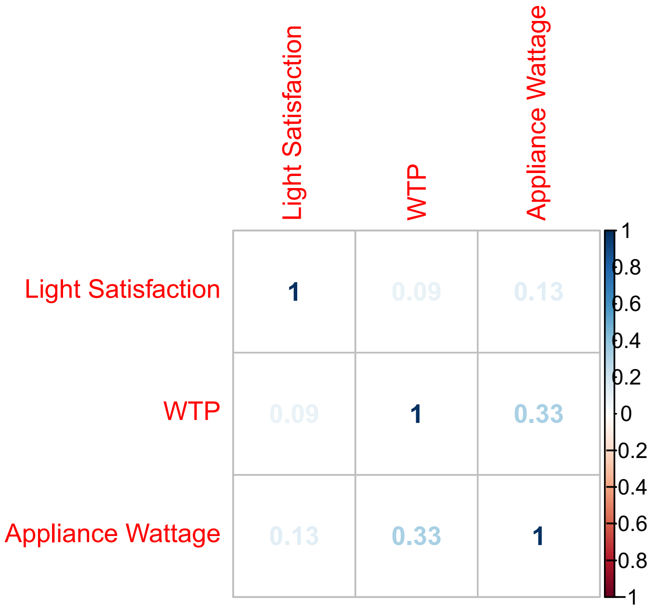

Fig. 1. This figure displays the correlation between our segmentation variables. Darker shades of blue suggest a higher positive correlation. Appliance wattage and willingness to pay are positively correlated, whereas the correlation between satisfaction with lighting and the other two variables is weak.

[2] In the analysis presented in the manuscript, we replace all missing values with zero. This approach implicitly assumes that households’ willingness to pay for electricity is 0. As a robustness check, we conduct all the analyses again dropping households that did not answer our willingness to pay questions. This analysis is available in the Appendix in Section S4.

[3] This question should capture the highest willingness to pay by respondents.

We also test the effect of several psychographic variables. In total, we employ six such variables which we collapse into three factors using latent variable factor analysis. The number of factors was determined using parallel analysis, which compares the eigenvalues of the correlation matrix of the observed data to that of a random dataset. The three factors capture leadership, cheap consumership, and risk aversion. Similar variables have shown strong associations with demand for particular types of electricity in past studies (Rowlands et al., 2003).

Finally, we include clearly endogenous variables to predict membership. These include household grid connection, household minigrid connection, and household expenses. Although the SPI interventions were not randomized and are certainly correlated with village characteristics, this analysis allows understanding into whether their customer base is correlated with any certain cluster membership. However, we are unable to say whether SPI interventions influenced membership or whether these customer characteristics predicted the choice of intervention locations, as they were not randomized. We are agnostic as to which direction causation runs.

Summary statistics for these variables, along with segmentation variables, are displayed in Table 1. All variables exhibit a high level of variation, although it is clear that there are some outliers with respect to appliance wattage. The measurement of variables is described in their corresponding sections.

In all models, we also include state fixed effects in order to account for unobserved heterogeneity across states.

4. Results

In this section, we present the results of our analysis. First, we determine the optimal number of clusters using the gap statistic method. The optimal number of clusters is determined to be three. We then create the three clusters using the *k*-means algorithm, in which the within-cluster difference between observations in terms of the chosen variables is minimized. Once we have determined the clusters, we characterize them based on their mean values. We then determine the predictors of cluster membership through regression analysis.

4.1. Determining the number of clusters

We determine the optimal number of clusters using the gap statistic, which compares the change in within-cluster dispersion with that under a reference null distribution (Tibshirani et al., 2001). Our approach is to use the firstSEmax process in the cluster package in R (Maechler et al., 2019). This approach " looks for the smallest *k* such that its value $f(k)$ is not more than 1 standard error away from the first local maximum" (Maechler et al., 2019). In our test, we limit the possible number of clusters to ten. We find the optimal number of clusters to be three, as shown in Fig. 2.

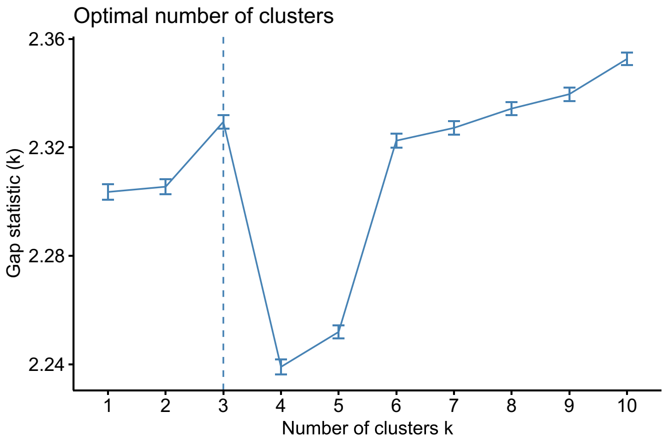

Fig. 2. The optimal number of clusters computed by the gap statistic. Three clusters is found to be the optimal number.

4.2. Cluster membership

To separate the data into three clusters, we use hard *k*-means cluster analysis, in which each observation is assigned to only exactly cluster. The approach minimizes the difference between observations within clusters while maximizing between-cluster heterogeneity. Before the clustering method was employed, we centered each variable by subtracting the variable means from each variable and dividing them by their standard deviation.

In Fig. 3, we show the scaled characteristics of each cluster along segmentation variables. The figure shows that the data is separated into three clusters with clear variation across the segmentation variables. Cluster 1 has the lowest values of all three variables, whereas cluster 3 has the highest values for willingness to pay and appliance wattage. Cluster 2 has middle values for willingness to pay and appliance wattage, and the highest mean value for satisfaction with lighting. We show the number of customers in each cluster in Section S2 in the Appendix. Cluster 2 is the largest, with 5337 observations. Cluster 1 has 4234 observations, whereas cluster 3 has 678.

Table 1
Summary statistics for segmentation and predictor variables for the entire sample. Missing data were imputed to be 0. $N = 10{,}249$.

Statistic	Mean	St. Dev.	Min	Max
Willingness to Pay	203.951	248.374	0	10,000
Satisfaction with Light	3.336	0.952	1	5
Appliance Wattage	325.693	578.866	0.000	9564.000
Age	43.521	14.491	18	95
Uneducated?	0.468	0.499	0	1
Schooling?	0.472	0.499	0	1
Higher Education?	0.059	0.236	0	1
Hindu?	0.881	0.324	0	1
Scheduled Caste or Tribe?	0.274	0.446	0	1
Monthly Household Expenses (log)	8.487	0.667	5.303	11.513
Household Size	6.470	3.376	1	45
Village Hours of Grid Electricity	11.420	5.932	0.000	21.960
Leader?	−0.000	0.868	−1.445	1.994
Cheap?	−0.000	0.998	−1.732	1.895
Risk Averse?	−0.000	0.639	−1.403	1.711
Grid Connection?	0.743	0.437	0	1
Mini-grid Connection?	0.020	0.139	0	1

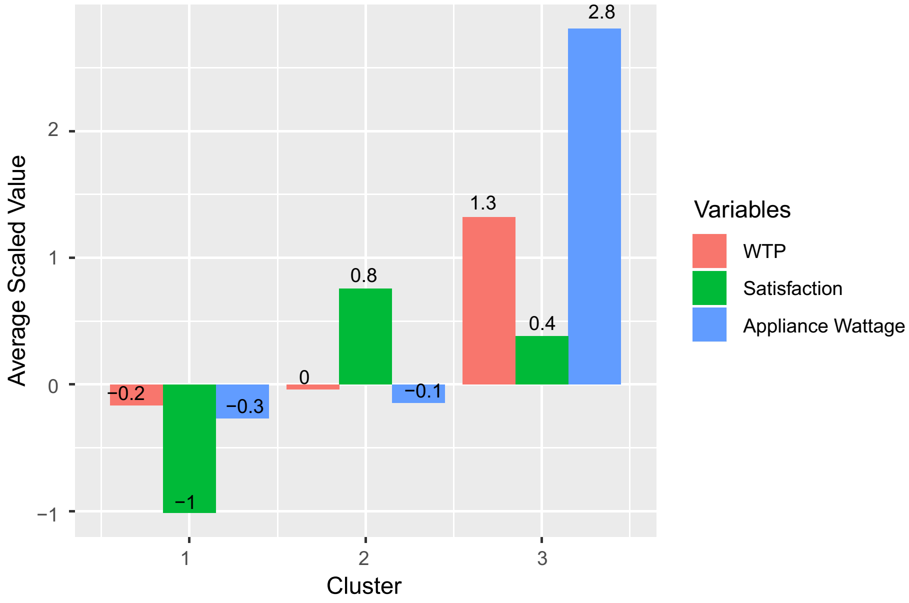

Fig. 3. Scaled mean values of variables used for segmentation by cluster. There is a high level of variation between clusters in the means of our segmentation variables. We characterize cluster 1 ($n = 4234$) as "potential customer", cluster 2 ($n = 5337$) as "low-demand" and cluster 3 ($n = 678$) as "high-use". Variables are centered and divided by their standard deviation. WTP means willingness to pay.

We further show the unscaled values by cluster in Fig. 4. This graphic demonstrates the large difference in mean values between clusters. Significantly, we show that cluster 2 is characterized by high levels of satisfaction with lighting, whereas cluster 3 is characterized by high levels of willingness to pay and appliance wattage. This finding highlights the value of the segmentation approach: we effectively isolate two important types of customers. We argue that customers in cluster 2 are likely to be takers, or consumers who adopt the technology provided but potentially do not demand higher levels of it. Cluster 3, meanwhile, exhibits high demand for electricity.

We also display bivariate plots of the clusters according to the segmentation variables in order to characterize the clusters. These plots are shown in Fig. 5. Note that there is high inter-cluster variation across all variables, and moreover the high-use cluster contains the majority of outlier variables.

Given the characteristics of the clusters, we assign them qualitative meaning to abstract away from their characteristics. To this end, we consider cluster 1 to be the "potential customer" cluster, cluster 2 to be the "low-demand cluster" and cluster 3 to be the "high-use" cluster. Summary statistics separated by cluster, and including the wattage of all appliances included in segmentation are shown in the Appendix Section S3.

Finally, we display the clusters by state and SPI intervention. The distribution of cluster membership by state is shown in Fig. 6. The proportion of observations in each cluster is fairly consistent across states. The distribution by intervention is shown in Fig. 7. We see a much larger proportion of low use observations in non-intervention villages rather than in mini-grid or distribution franchise villages.

5. Predictors of cluster membership

Given that we can characterize three clusters in our dataset, we can then employ other variables to determine which predict membership in the three clusters. To do so, we use OLS regression, employing the predictor variables outlined above. In total, we estimate nine models. each cluster membership is coded as a binary dependent variable, and for each we estimate three models varying the inclusion of covariates and endogenous variables. For all models, we include state fixed effects and cluster standard errors at the village level. In each model, coefficients can be interpreted as the increase in probability of being in one cluster versus all others for a

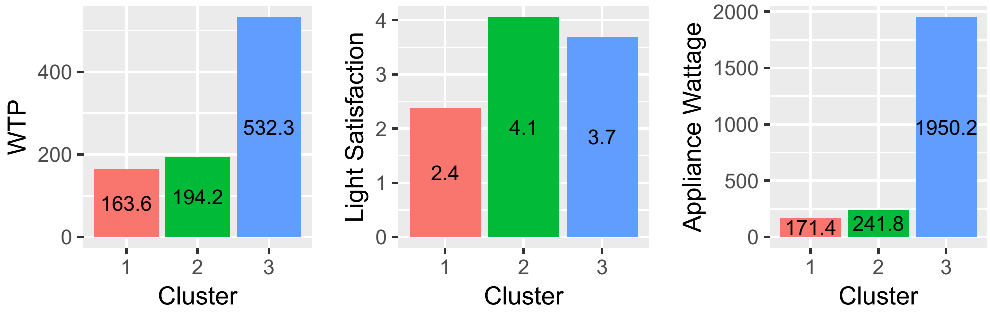

Fig. 4. Unscaled values of our segmentation variables. Cluster 3 has the highest values for willingness to pay and total appliance wattage, whereas cluster 2 has the highest satisfaction with lighting. WTP means willingness to pay.

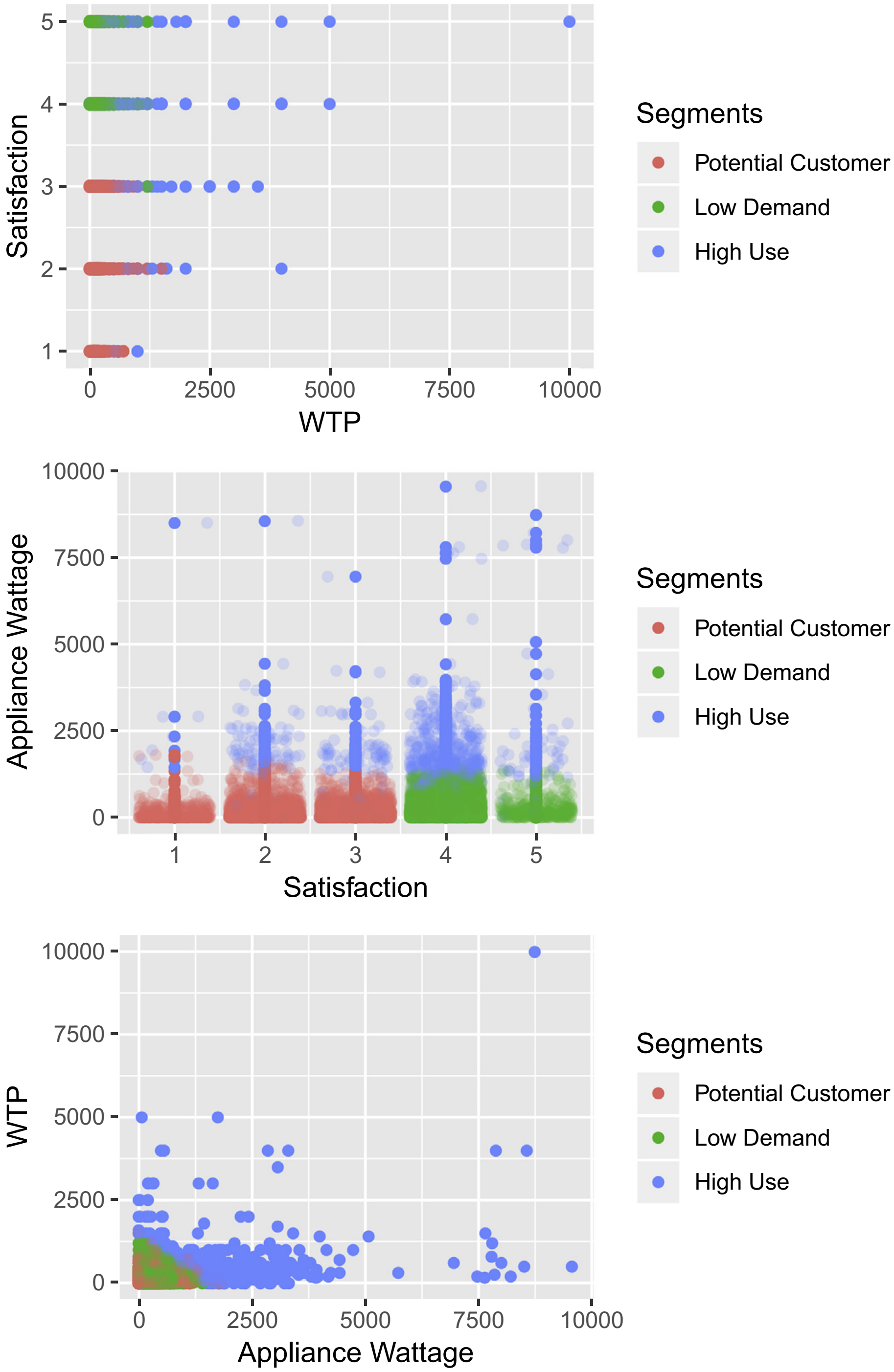

Fig. 5. Bivariate plots of segmentation variables by cluster. The plots exhibit high inter-cluster variation along all variables. The high-use cluster contains the majority of outlier variables. WTP means willingness to pay.

one unit change in the independent variable. Our general estimation equation is

$$Y_{ivs} = \beta X_i + \mu_s + \epsilon_v$$

where Y_i is membership in a cluster at the individual level, X_i is a vector of covariates for individual i, μ_s are state fixed effects, and ϵ_v is the error term clustered at the village (v) level. In Appendix Section S5, we also employ a multinomial logit model including all covariates and endogenous variables and plot the effects of each of the variables on cluster membership. The effects plots confirm the results shown in the linear probability models.

We begin by determining the predictors of membership in the potential customer cluster. Our first model includes only plausibly exogenous variables at the individual level in order to determine their effects without possible confounding. These are socioeconomic variables, such as age, education level, religion, caste, and household size. We then include psychographic factor variables in the second model. The third model adds endogenous variables to the model,

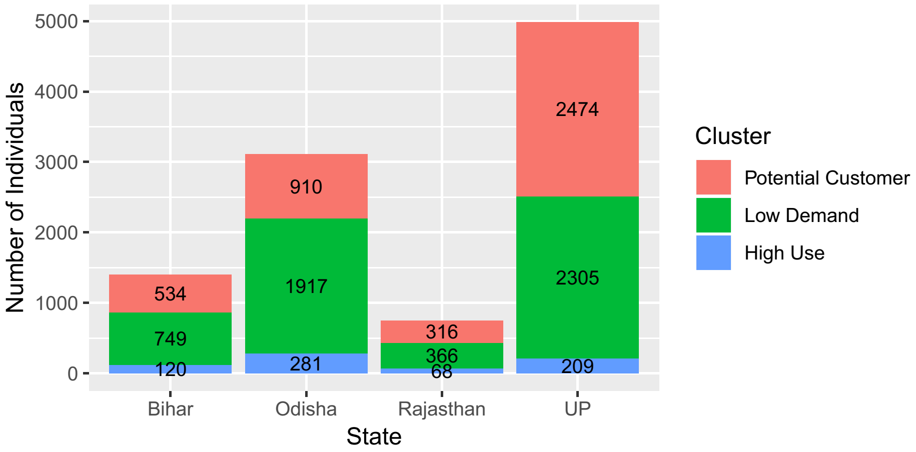

Fig. 6. Cluster membership by State. The distribution of observations in each cluster appears consistent across all States, although Rajasthan has fewer members in the high-use cluster.

including the mean hours of grid electricity available in a village, and whether households use grid or mini-grid electricity. The results of these models for the potential customer cluster are shown in Table 2.

In terms of socioeconomic variables, we find that age, and both levels of education tested have a significantly negative effect on the probability of membership in the potential customer cluster across models. Conversely, being a member of a scheduled caste or tribe increases the likelihood that observations are in the potential customer cluster. Household size does not have a significant coefficient except when endogenous variables are included in the model.

Leadership traits and risk aversion both decrease the likelihood of being in the potential customer cluster. This could be because these individuals already have sufficient access to energy. However, the effects of preferring cheap goods are weak, demonstrating that energy demand may be seen as a necessary good, and not one subject to budget considerations.

Finally, all endogenous variables have significant effects, although these results should be interpreted cautiously. Households with greater monthly expenses are less likely to be in the potential customer cluster. The same can be said for villages with greater hours of grid electricity and households that have grid and mini-grid connections. These results are not surprising, as energy supply should already be sufficient in these areas.

We then estimate the same models for membership in the low-demand cluster. These results are shown in Table 3. As the age of respondents increases, so does their likelihood of being in the low-demand cluster. However, being in a scheduled caste or tribe and having a larger household decreases the likelihood of being in this cluster. Other exogenous socioeconomic covariates are not statistically significant. This indicates that households with potential demand are generally those who are younger, and in a lower socioeconomic class.

Among the three psychographic variables, leadership traits and preferences for cheap goods predict membership in this segment. Risk aversion does not have predictive power.

In the third model, the endogenous variables all have significant coefficients. An increase in household expenses decreases the likelihood of being in this cluster. Conversely, having a grid or mini-grid connection increases the likelihood of membership in the low-demand cluster.

Finally, we estimate the three models for the high-use cluster. Results are shown in Table 4. In Model 1, we find that having attended school, and having acquired a higher education increase the likelihood of being in this cluster. Meanwhile, being in a scheduled caste or tribe lower the likelihood. Having a larger household also significantly increases the likelihood of being in the high-use cluster.

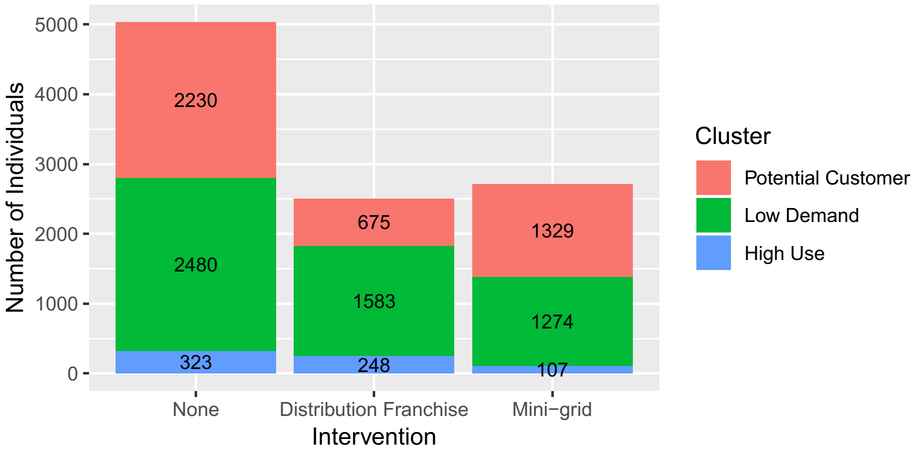

Fig. 7. Observations in each cluster by Smart Power India intervention. More observations are in the potential customer cluster in villages without SPI interventions.

Table 2
In this table, we present results for the predictors of membership in the "potential customer" cluster. We increase the number of predictors included in each model. Coefficients can be interpreted as the increased likelihood of being in the cluster given a one-unit increase in the independent variable. We show significant effects for several variables which are interpreted in the text.

	Dependent variable:		
	Potential Customer		
	(1)	(2)	(3)
Age	−0.001**	−0.001**	−0.001**
	(0.0004)	(0.0004)	(0.0004)
Schooling?	−0.053***	−0.045***	−0.029***
	(0.011)	(0.011)	(0.011)
Higher Education?	−0.138***	−0.116***	−0.075***
	(0.020)	(0.020)	(0.020)
Hindu?	−0.035	−0.024	−0.029
	(0.024)	(0.023)	(0.023)
Scheduled Caste/Tribe?	0.067***	0.065***	0.033$*$
	(0.019)	(0.019)	(0.018)
Log Expenses			−0.044***
			(0.011)
Household Size	−0.003	−0.002	0.003$*$
	(0.002)	(0.002)	(0.002)
Village Grid Hours			−0.015***
			(0.003)
Leader		−0.082***	−0.078***
		(0.010)	(0.010)
Cheap		−0.008	−0.009
		(0.006)	(0.006)
Risk Averse		−0.028**	−0.026**
		(0.011)	(0.011)
Grid Connection?			−0.173***
			(0.021)
Mini-Grid Connection?			−0.264***
			(0.048)
Constant	0.522***	0.532***	1.268***
	(0.062)	(0.061)	(0.117)
State Fixed Effects?	Yes	Yes	Yes
Observations	10,249	10,249	10,168
Adjusted R^2	0.043	0.063	0.117

Note: $*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.
Standard errors clustered at village level.

Table 3
This table displays results using membership in the "low-demand" cluster as the dependent variable. We increase the number of predictors included in each model. Coefficients can be interpreted as the increased likelihood of being in the cluster given a one-unit increase in the independent variable. We show significant effects for several variables which are interpreted in the text.

	Dependent variable:		
	Low demand		
	(1)	(2)	(3)
Age	0.001**	0.001**	0.001**
	(0.0004)	(0.0004)	(0.0004)
Schooling?	0.012	0.009	0.006
	(0.011)	(0.011)	(0.011)
Higher Education?	−0.005	−0.014	−0.022
	(0.025)	(0.024)	(0.025)
Hindu?	0.026	0.016	0.018
	(0.024)	(0.023)	(0.024)
Scheduled Caste/Tribe?	−0.034$*$	−0.034$*$	−0.016
	(0.019)	(0.019)	(0.018)
Log Expenses			−0.026**
			(0.012)
Household Size	−0.003$*$	−0.004**	−0.004$*$
	(0.002)	(0.002)	(0.002)
Village Grid Hours			0.011***
			(0.003)
Leader		0.066***	0.062***
		(0.010)	(0.011)
Cheap		0.029***	0.028***
		(0.007)	(0.007)
Risk Averse		0.009	0.008
		(0.011)	(0.011)
Grid Connection?			0.148***
			(0.022)
Mini-Grid Connection?			0.262***
			(0.049)
Constant	0.455***	0.441***	0.363***
	(0.057)	(0.056)	(0.120)
State Fixed Effects?	Yes	Yes	Yes
Observations	10,249	10,249	10,168
Adjusted R^2	0.020	0.036	0.065

Note: $*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.
Standard errors clustered at village level.

All three psychographic variables have significant effects. Having leadership traits and being risk averse increases the likelihood of being in the high-use cluster. Preferring cheap goods increases the likelihood of not being in the high-use cluster. This indicates that some behavioral measures influence demand, although the mechanism is unclear.

Finally, in Model 3, three endogenous variables have positive effects. Increases in household expenses, hours of village grid electricity, and household grid connections increase the likelihood of being in the high-use cluster.[4] However, mini-grid connections surprisingly do not.

Overall, we show that observable characteristics of individuals can predict their cluster membership and thus their projected level of demand. This analysis can be used to develop more effective targeting and messaging strategies for organizations attempting to increase the uptake of electricity. For example, our results suggest that in places with a large number of customers exhibiting characteristics similar to those in the potential customer segment, a focus on increasing the quality of electricity service can greatly increase demand.

6. Conclusion and policy implications

The above analysis highlights the role that easily observable covariates play in predicting customers' current and future demand for electricity. This indicates that organizations seeking to advertise and promote the uptake of different electricity technologies in rural India can vary their campaigns and strategies along socioeconomic lines. Moreover, we provide evidence that psychographic variables predict electricity consumption as well, meaning that behavioral trends within a community may predict their willingness to use new electricity products. For instance, individuals who have leadership traits and exhibit risk aversion may be more willing to use electricity, whereas those who do not may not be effective customers to target.

We also show that the availability of electricity is a strong predictor of being in a high-use cluster. Such a relationship is tenuous, of course, because electricity may be available only in those areas where we would expect high demand. However, even when analysis is conducted at the village level, more availability leads to a higher likelihood of being in the high-use cluster. This indicates that simply expanding availability may increase demand.

The generation of the segments itself also contributes to our understanding of the electricity market in rural India. First, we show that segmentation is easy to accomplish given even a limited dataset. We create three clusters of individuals using data on willingness to pay, appliance usage, and satisfaction with lighting. Similar analysis could be conducted on other pre-existing datasets. Second, we show

[4] It is important to note that the coefficient for grid connection is negative in Table S4 of the Appendix, where missing data for willingness to pay were dropped instead of treated as 0s.

Table 4
This table displays results using membership in the "high-use" cluster as the dependent variable. We increase the number of predictors included in each model. Coefficients can be interpreted as the increased likelihood of being in the cluster given a one-unit increase in the independent variable. We show significant effects for several variables which are interpreted in the text.

	Dependent variable:		
	High use		
	(1)	(2)	(3)
Age	0.0001	0.0001	0.0001
	(0.0002)	(0.0002)	(0.0002)
Schooling?	0.041***	0.036***	0.023***
	(0.005)	(0.005)	(0.005)
Higher Education?	0.143***	0.130***	0.096***
	(0.019)	(0.019)	(0.018)
Hindu?	0.009	0.007	0.011
	(0.009)	(0.009)	(0.008)
Scheduled Caste/Tribe?	−0.033***	−0.031***	−0.017***
	(0.006)	(0.006)	(0.006)
Log Expenses			0.070***
			(0.006)
Household Size	0.006***	0.005***	0.0004
	(0.001)	(0.001)	(0.001)
Village Grid Hours			0.005***
			(0.001)
Leader		0.016***	0.016***
		(0.003)	(0.003)
Cheap		−0.021***	−0.019***
		(0.004)	(0.003)
Risk Averse		0.019***	0.018***
		(0.005)	(0.004)
Grid Connection?			0.025***
			(0.005)
Mini-Grid Connection?			0.002
			(0.011)
Constant	0.023	0.027	−0.631***
	(0.019)	(0.019)	(0.058)
State Fixed Effects?	Yes	Yes	Yes
Observations	10,249	10,249	10,168
Adjusted R^2	0.041	0.050	0.091

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.
Standard errors clustered at village level.

that creating a small number of segments can still create clusters with higher levels of inter-cluster variation along our segmentation variables. This indicates that standard measures of existing or potential demand for electricity are correlated and that clear consumer groups can be created along these lines. Such an approach to understanding and describing the electricity market may be fruitful in other under-electrified areas.

The segmentation exercise undertaken in this paper contributes to our understanding of the rural Indian electricity market in two key ways. First, it describes the rural electricity market along three key lines: willingness to pay, satisfaction with lighting and appliance usage. We show that three clusters are optimal for segmenting along these variables, and create potential customer, low-demand and high-use clusters. Second, it determines the predictors of cluster membership. We demonstrate that socioeconomic, psychographic, and electricity availability variables all affect the likelihood of a given household falling within a certain usage cluster.

The analysis in this essay is by no means exhaustive or conclusive, as segmentation is inherently an exercise defined by a large amount of researcher degrees of freedom. If we began with different segmentation variables, our optimal number of clusters may have differed, as would have the effect of our predictor variables. Moreover, we focus only on demand for electricity in general and not on demand for specific services. Therefore, our analysis is limited in speaking to customer segments in specific service markets. This would certainly be a fruitful avenue for future research. However, we believe the findings are an effective step towards understanding the composition of the electricity market in rural India. Moreover, this study can contribute to future research on electricity and technological uptake in general. For example, it indicates that in future RCTs examining electricity uptake, attention should be paid to heterogenous responses by cluster.

CRediT authorship contribution statement

Daniel Robert Thomas: Methodology, Investigation, Software, Writing - original draft, Writing - review & editing. **Shalu Agrawal:** Investigation, Writing - original draft, Writing - review & editing. **S.P. Harish:** Investigation, Writing - original draft, Writing - review & editing. **Aseem Mahajan:** Investigation, Writing - original draft, Writing - review & editing. **Johannes Urpelainen:** Conceptualization, Project administration, Writing - original draft, Writing - review & editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eneco.2020.104697.

References

Aklin, M., Bayer, P., Harish, S.P., Urpelainen, J., 2018. Economics of household technology adoption in developing countries: evidence from solar technology adoption in rural India. Energy Econ. 72, 35–46.

Assael, H., Roscoe, A.M., 1976. Approaches to market segmentation analysis. J. Mark. 67–76.

Bhattacharyya, S.C., 2006. Renewable energies and the poor: niche or nexus? Energy Policy 34 (6), 659–663.

Dutta, G., Mitra, K., 2015. Dynamic Pricing of Electricity: A Survey of Related Research.

Foedermayr, E.K., Diamantopoulos, A., 2008. Market segmentation in practice: review of empirical studies, methodological assessment, and agenda for future research. J. Strateg. Mark. 16 (3), 223–265.

Gaur, K., Harish, K.R., Agarwal, P.K., Baba, K.V.S., Soonee, S.K., 2016. Analysing the Electricity Demand Pattern. Power System Operation Corporation Limited, New Delhi, India.

Homburg, C., Koschate, N., Hoyer, W.D., 2005. Do satisfied customers really pay more? A study of the relationship between customer satisfaction and willingness to pay. J. Mark. 69 (2), 84–96.

Hyland, M., Leahy, E., Tol, R.S.J., 2013. The potential for segmentation of the retail market for electricity in Ireland. Energy Policy 61, 349–359.

Jones, R.V., Fuertes, A., Lomas, K.J., 2015. Socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings. Renew. Sustain. Energy Rev. 43, 901–917.

Kemmler, A., 2007. Factors influencing household access to electricity in India. Energy Sustain. Dev. 11 (4), 13–20.

Kiran Mor, S., 2014. Segmenting Indian consumers: a psychographic approach. Glob. J. Manag. Bus. Res. 14 (3), 33–43.

Lewis, J.J., Pattanayak, S.K., 2012. Who adopts improved fuels and cookstoves? A systematic review. Environ. Health Perspect. 120 (5), 637.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2019. cluster: Cluster Analysis Basics and Extensions.. R package version 2.1.0 — For new features, see the 'Changelog' file (in the package source).

Narang, R., 2011. Examining the role of various psychographic characteristics in apparel store selection: a study on Indian youth. Young Consum. 12 (2), 133–144.

Oda, H., Tsujita, Y., 2011. The determinants of rural electrification: the case of Bihar, India. Energy Policy 39 (6), 3086–3095.

Palit, D., Chaurey, A., 2011. Off-grid rural electrification experiences from South Asia: status and best practices. Energy Sustain. Dev. 15 (3), 266–276.

Rajabi, A., Li, L., Zhang, J., Zhu, J., Ghavidel, S., Ghadi, M.J., 2017. A review on clustering of residential electricity customers and its applications. Electrical Machines and Systems (ICEMS), 2017 20th International Conference on. IEEE., pp. 1–6.

Rowlands, I.H., Scott, D., Parker, P., 2003. Consumers and green electricity: profiling potential purchasers. Bus. Strateg. Environ. 12 (1), 36–48.

Singh, A.K., 2007. Rural Marketing: Indian Perspective. New Age International.

Smith, M.G., Urpelainen, J., 2014. Early adopters of solar panels in developing countries: evidence from Tanzania. Rev. Policy Res. 31 (1), 17–37.

Srinivasan, S., 2005. Segmentation of the Indian photovoltaic market. Renew. Sustain. Energy Rev. 9 (2), 215–227.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Stat Methodol.) 63 (2), 411–423.

Wedel, M., Kamakura, W.A., 2012. Market Segmentation: Conceptual and Methodological Foundations. vol.8. Springer Science & Business Media.

Zhang, L., Wu, Y., 2012. Market segmentation and willingness to pay for green electricity among urban residents in China: the case of Jiangsu Province. Energy Policy 51, 514–523.