Proceedings of the ASME 2012 International Design Engineering Technical Conferences &
Computers and Information in Engineering Conference
IDETC/CIE 2012
August 12-15, 2012, Chicago, IL, USA

DETC2012-70468

AN APPROACH FOR REVEALED CONSUMER PREFERENCES FOR TECHNOLOGY PRODUCTS: A CASE STUDY OF RESIDENTIAL SOLAR PANELS

Heidi Q. Chen
Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
Email: heidiqc@mit.edu

Tomonori Honda
Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
Email: tomonori@mit.edu

Maria C. Yang[*]
Mechanical Engineering and
Engineering Systems
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
Email: mcyang@mit.edu

ABSTRACT

Consumer preferences can serve as an effective basis for determining key product attributes necessary for market success, allowing firms to optimally allocate time and resources toward the development of these critical attributes. However, identification of consumer preferences can be challenging, particularly for technology-push products that are still early on in the technology diffusion S-curve, which need an additional push to appeal to the early majority. This paper presents a method for revealing preferences from actual market data and technical specifications. The approach is explored using three machine learning methods: Artificial Neural Networks, Random Forest decision trees, and Gradient Boosted regression applied on the residential photovoltaic panel industry in California, USA. Residential solar photovoltaic installation data over a period of 5 years from 2007-2011 obtained from the California Solar Initiative is analyzed, and 3 critical attributes are extracted from a pool of 34 technical attributes obtained from panel specification sheets. The work shows that machine learning methods, when used carefully, can be an inexpensive and effective method of revealing consumer preferences and guiding design priorities.

NOMENCLATURE

k	Attribute number from 1-34
MSE	Mean Squared Error
PTC	PV-USA Test Conditions: Air temperature 20°C, 10m above ground level, 1m/s wind speed, AM1.5 solar spectrum, 1000W/m^2 irradiance.
R	Correlation coefficient
R^2	Coefficient of determination
STC	Standard Test Conditions: Cell temperature 25°C, AM1.5 solar spectrum, 1000W/m^2 irradiance.
σ_{MS}	Standard deviation of market share

INTRODUCTION

Within a firm there is a constant need to rise above the competition and gain market success. In order to achieve this goal, firms are constantly trying to find ways to appeal to the customer by determining their changing needs, wants, values and behavior and designing for them [1]. However, in the fast paced world of product development, there is a limit on the time and resources that can be allocated to product development. Hence, the identification of key product attributes that contribute to a product's market success is crucial, especially in the early stages of product development where the conceptual design phase can account for a large percentage of the overall manufacturing cost [2]. It is in the interest of both designers and stakeholders to know how to optimally allocate resources in order to increase the likelihood of market success.

This is especially so for technology-push products that are still within the early part of the technology diffusion S-curve,

[*] Address all correspondence to this author.

where there is slow uptake of the technology and product features have not fully matured [3]. Since the technology is still considered new, only early adopters have warmed to the product, and there is potential for much feature and market expansion. Knowing what the customer wants at this stage is necessary for the product to bridge the chasm between the early adopters and the early majority, allowing the product to flourish in the market [4].

This paper proposes an approach to extracting consumer preferences by determining critical attributes using the established revealed preference framework [5–8], and drawing on advances in computational intelligence and machine learning to support the analysis. Revealed preference methods have been used widely in economics research, but little has been done in the area of design applications. The main research questions are:

1. Can revealed consumer preferences be obtained from market data and engineering specifications using machine learning methods?
2. Is there agreement among the machine learning methods that suggest the validity of the data and methods?

We present a case study of residential solar photovoltaic panels in the California market to illustrate our methodology. Engineering specification data obtained from solar panel data sheets combined with real market data available freely on the California Solar Initiative database is analyzed using 3 machine learning methods to quickly assess critical technical attributes from the dataset.

BACKGROUND

Much work has been done within the academic community to determine consumer preferences using choice modeling. These can be broken down into 2 main categories: stated preference methods which measure consumers' explicit preferences over hypothetical alternatives, and revealed preference methods which extract preferences from actual market data [9].

Over the years, stated preference methods have gained ground in the marketing community due to their flexibility and ease of implementation. Popular survey based stated preference methods include self-explicated methods like Kelly's repertory grid [10, 11], Self-Explicated Method (SEM) [12] and the Single Unit Marketing Model [13] among others, requesting consumers to rank or rate various product attributes. Another group of stated preference methods where relative preferences are obtained include MaxDiff [14], and conjoint analysis [15, 16], which ask consumers to choose between different products which have varying attributes. Multiple hybrid models that incorporate both self-explicated and relative preferences also exist. Non-survey based methods include focus groups and field observations, which require considerable time, expertise and resources to carry out, and may be hard to quantify.

The potential problem with these stated preference methods is that consumers often exhibit preference inconsistencies, constructing their preferences along the way, or changing their preferences due to some shift in the phrasing of the questions [17]. Research on the accuracy of consumers' predictions show a disconnect between preferences obtained during preference elicitation and actual decision making [18]. Stated preference methods have also come under considerable criticism because of the belief that consumers react differently under hypothetical experiments compared to when they are faced with the real market situation [19, 20].

In comparison, revealed preference methods could be a better reflection of purchase behavior than stated preference methods as they take into account external factors like third party influences that might affect the consumer's decision. This has been expressed in the economics and decision making literature to be especially important if the consumer's choice is based heavily on the recommendation of a more experienced expert, as a result of complexity inherent in the product, or limited personal experience [21]. However, revealed preference methods have been difficult to implement due to several factors. These include the high cost of collecting large sets of relevant data, limited technological knowledge, problems with multicollinearity, and the inability to test new variables [22]. As technology has improved and computer processing has become increasingly fast, efficient and cost effective, it has become feasible to reevaluate these methods. Furthermore, more companies are keeping digital records of product sales, making data collection less of a burden than before. Machine learning methods that are capable of dealing with multicollinearity involving regression and classification can now be applied on large sets of marketing data, overcoming the issue with multicollinearity that several academics have identified, and allowing for the identification of key attributes in an efficient way. Finally, the inability to test new variables still poses a significant challenge, as the new variables may be outside the data range, and involve extrapolation outside the range used to create the model. This can be dealt with by a careful use of stated preference methods in combination with the revealed preference framework, which the authors of this paper are working toward.

Similar work that has been done in the joint field of product design and machine learning include: Agard and Kunsiak's work on data mining for the design of product families [23], where algorithms were used for customer segregation; Ferguson et al's work on creating a decision support system for providing information from later to earlier stages in the design process [24]. A good overview of other applications of computational intelligence in product design engineering can be found in Kusiak's 2007 review [25].

This paper sets itself apart in the design community by focusing on revealed preferences instead of stated preferences as a means to extract consumer purchasing preferences. Compared to existing data mining methods, we take data from widely available sources instead of from within the company, combining real market data and engineering specifications from data sheets in order to determine a set of critical attributes that can be prioritized to boost a firm's competitiveness. Furthermore, the focus is on finding key attributes that impact the design decision instead of predicting market share. Lastly, the result of machine learning algorithms are compared to validate their effectiveness.

CASE STUDY: RESIDENTIAL SOLAR PHOTOVOLTAIC (PV) SYSTEMS

In recent years, the US government has been encouraging the development of renewable power, providing the solar industry with increased funding for the development of solar panels for residential, commercial and utility deployments. Residential installations in particular have gained attention due to their use of otherwise "dead" space, utilizing area on rooftops or facades for the panels. Generous subsidies and rebates have been put into place in order to encourage homeowners to adopt the technology. Despite this, the industry is still considered by many to be in the early stages of the technology diffusion S-curve, with few homeowners choosing to purchase PV systems for their properties.

This paper proposes the view of considering solar panels as a product rather than a technology. Products differ from technology in that they may be described by both qualitative and quantitative characteristics, and are designed to appeal to consumers. Much of the current academic engineering focus on solar panels has rightly been on the science and technology behind the application, improving the performance of the conversion of sunlight to electricity and increasing the reliability and durability of the system. This is critical for spurring increases in the demand for large scale facilities installations. At the same time, it is important to convince consumers to purchase a PV system at the residential level where decision makers are spread out and individual households have different requirements. There is limited academic literature on understanding consumer needs in order to increase adoption. Existing research is centered on identifying characteristics of adopters [26], discovering their motives for acquiring a PV system [27], determining barriers to adoption [28], and understanding the link between attractive factors of PV systems [29]. However, these studies are limited to stated preference studies, and do not include real market data or technical specifications.

At this moment, the industry is also facing an oversupply condition, a result of an increase in global manufacturing with little corresponding increase in the demand on the side of the consumer. Especially now that the survival of companies are at stake, funding is tight and profits are low, there is a need to focus available resources on high priority features that will lead to more consumer purchasing to increase the firm's market share and maintain profits. The state of the industry thus lends itself to our study.

For this paper, we make use of market share as a reflection of market success, even though the definition of market success varies widely in literature [30]. Market share was chosen as it is publicly available, unlike customer satisfaction levels, revenue or profits which are usually kept within the company and are difficult or costly to measure. It has also been discovered to be the most useful customer-based measure for the success of line extensions of existing products [31]. The aim of this project is to see if there is a correlation between photovoltaic panel technical specifications, and their success in the market, measured by market share. In this way, designers will be able to better optimize design priorities that lead to product success.

Technical Features of Residential PV Panels in the California Market

The working data set published in September 7, 2011 from the California Solar Statistics California Solar Initiative incentive application database [32] served as the paper's source of market data. The data is considered representative of the USA solar consumption, as California is the current leading producer of solar power in the United States, accounting for 44% of the total grid-connected PV cumulative installed capacity through quarter 3 of 2011 [33]. The working database includes all incentive applications from January 2007 to November 2011 made in California, hence includes both successful subsidized installations and unsuccessful incentive applications made by a variety of consumers. It was assumed that unsuccessful incentive applications did not result in a PV installation.

The data was filtered to include only residential installations with a completed installation status, excluding applications that are from the commercial, government or non-profit system owner sector, as well as those that were canceled or pending. This was done in order to concentrate on the small scale PV systems that were actually installed during the 2007-2011 time-frame. Installations with more than 1 PV module type were filtered out, as the effective cost calculations cannot be done. Finally, new panels introduced during the past year were removed, as they are still too new and the market has not had adequate time to respond. After filtering, the data set was reduced from 73,514 to 32,896 installed systems with a total of 586 panel types, mostly due to filtering out non-residential systems. Filtering out systems with more than 1 PV panel type accounted for less than 0.8% of the total number of systems, and the effect of neglecting them in the subsequent calculations was taken to be negligible.

From this dataset, the quantity installed of each panel was calculated as a proxy for market share, and the panels ranked by that metric. Since a large portion of the market is controlled by a small subset of the 586 panels, as shown in Fig. 1, further analysis was required to find a cutoff point to focus the further analysis on the panels that are considered the most successful in the open California market. An established binary linear classi-

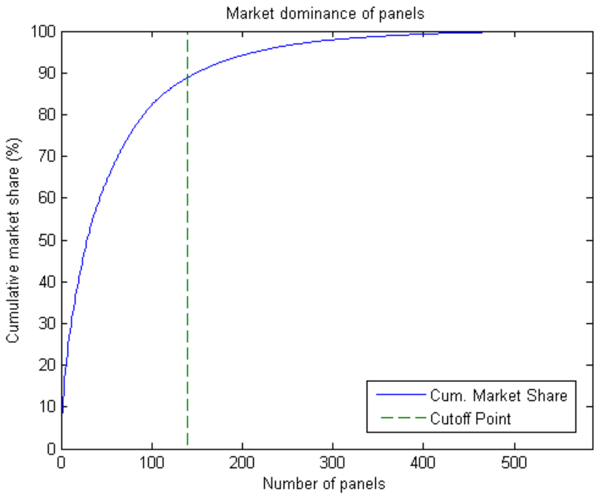

FIGURE 1. CUMULATIVE MARKET SHARE OF PANELS

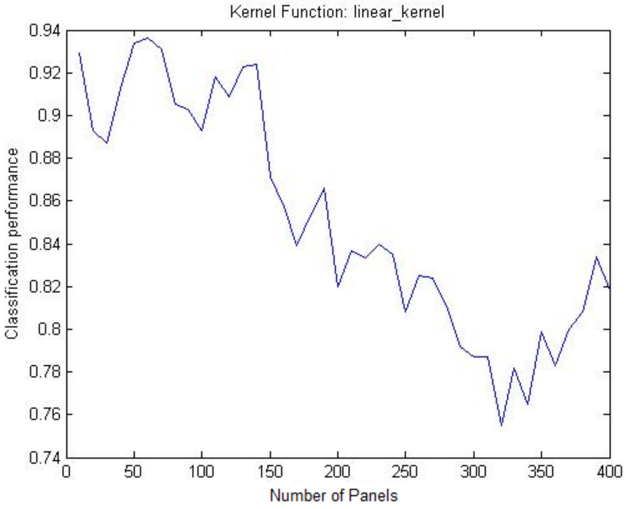

FIGURE 2. SVM CLASSIFICATION PERFORMANCE FOR CHOICE OF CUTOFF POINT

fier called Support Vector Machine (SVM) [34], was chosen. 200 panels picked at random were sourced for their technical specifications from manufacturer data sheets to form a specifications dataset. This dataset included common markers for technical performance, including attributes like efficiency tested at Standard Test Conditions (STC), rated power, and power warranty. These 22 markers were chosen as initial distinguishing attributes of the technical performance of each panel, as shown in Tab. 1A.

SVM took the set of PV panel data and categorized the panels into 2 groups by multiple attributes, including the 22 attributes stated in Table 1A and the panel's market share. A linear kernel was applied as it best suited the data spread. Figure 2 shows a noticeable drop-off in SVM classification performance at the top 140 panels, so that was chosen to be the cutoff point. This subset was determined to control 88.9% of the California market.

In the same way as the previous step, the 140 panels with the highest market share were identified by their model number and sourced for their technical specifications. From a combination of panel datasheets and marketing material, an expanded list of 34 attributes was identified (Tab 1B). This expanded list adds distinguishing characteristics of the panels, like appearance, packaging and environmental characteristics to the initial 22 attribute list, and is a more comprehensive collection of technical attributes.

As expected, the expanded list of attributes exhibited a high degree of multicollinearity, meaning that the attributes were highly correlated. This is a problem as it decreases the accuracy of the model. To reduce parameter correlation between the attributes and improve the multiple regression model, the redundant attributes were identified using a variance inflation factor (VIF) calculation, which quantifies the severity of multicollinearity in an ordinary least squares regression analysis. This method was chosen because of the ease of comparing multicollinearity between attributes. The variance inflation factor (VIF) for each attribute was calculated using Eqn. 1 by holding it as the dependent variable and performing regression with the rest of the attributes as independent variables.

$$VIF_k = \frac{1}{1-R_k^2}, \quad R_k^2 = 1 - \frac{MSE_k}{\sigma_k^2} \tag{1}$$

where k is the attribute number from 1-34. Attributes with high VIF values of more than 20 were removed from the specifications list [35], as shown in Tab. 1C. A total of 8 attributes were removed, leading to a reduced list of 26 attributes.

METHODOLOGY

Critical Attribute Determination

An overview of the methodology is presented in Fig. 3. A set of 3 computational machine learning regression methods were used to determine the important technical attributes that most influence market share. These methods were chosen over others as they are known in the machine learning community to be robust. However, other methods like SVM regression and Elastic Nets could have been used to achieve the same purpose.

1. *Artificial Neural Network (ANN) regression*
   ANN regression is a non-linear statistical data modeling that models complex relationships between inputs and outputs in a network of synapses and neurons. [36].

TABLE 1. ATTRBUTE DEFINITION LIST. (A) INITIAL SPECS USED FOR SVM ANALYSIS (B) EXPANDED SPECS USED FOR VIF ANALYSIS (C) FINAL REDUCED SPECS USED FOR 3 REGRESSION METHODS

Properties	**Specifications**	**A- SVM**	**B-VIF**	**C- Reg.**	**Definition**
Electrical	Rated power (W)	x	x		Power output of panel at STC
	Power variance (-%)	x	x	x	Negative power output tolerance
	Power variance (+%)	x	x	x	Positive power output tolerance
	Power at PTC (W)	x	x		Power output of panel at PTC
	Power ratio: PTC/STC	x	x	x	Power output ratio: PTC/STC
	Efficiency at STC (%)	x	x	x	Panel efficiency at STC
	Fill factor at STC		x	x	Fill factor of panel at STC
Physical	Weight (kg)	x	x		Total weight of panel
	Weight per W (kg/ W)	x	x	x	Weight of panel per Watt of rated power output
	Area of panel (m^2)	x	x		Area of panel
	Cell Number	x	x	x	Number of PV cells in panel
	Frame color (black/ not black)	x	x	x	Color of panel frame
	Thickness (mm)	x	x	x	Thickness of panel
	Length (mm)		x	x	Length of panel
	Width (mm)		x	x	Width of panel
	Appearance (even/ uneven)		x	x	Visual surface evenness of panel
	Cardboard free packaging		x		Panel packaging contains no cardboard
	Optimized packaging		x	x	Panel packaging optimized for least waste
	Lead-free		x	x	Solder used in panel is lead-free
	Tile		x		Panel in form of roof tiling
Certifications	IEC 61215 / IEC 61646	x	x	x	IEC PV design qualifcation and type approval
	IEC 61730	x	x	x	IEC PV module safety qualification
	UL 1703	x	x	x	UL Standard for safety of flat-plate PV panels
	CE Marking	x	x	x	Compliance with European conformity requirements
	IS0 9001	x	x	x	IS0 Quality management standard
	IS0 14001	x	x	x	IS0 Environmental management standard
	NEC 2008		x	x	NEC Safe installation of electrical equipment standard
	Safety class II @ 1000V		x	x	Double insulated appliance standard
	IEC 61701		x		IEC PV Salt mist corrosion standard
	UL 4703		x		UL PV cable standard
Warranty	Workmanship Warranty (years)	x	x	x	Workmanship warranty
	Power warranty (% power warranted years)	x	x	x	Power warranty, calculated for comparison by taking area of the % warrented by years warranted curve
Economics	Effective Cost/W ($/W)	x	x	x	Post subsidy system cost per Watt of rated power output
	Time on market (years)	x	x	x	Length of time panel has been on the market

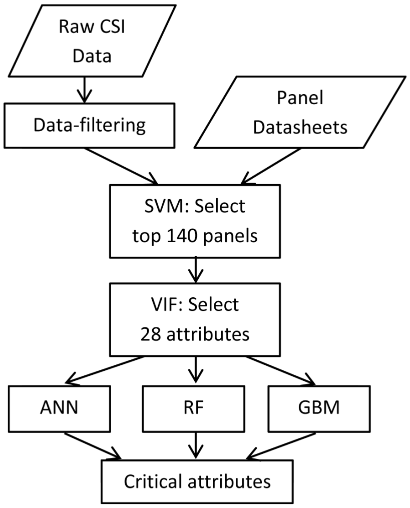

FIGURE 3. FLOWCHART OF METHODOLOGY

2. *Random Forest regression*
   Random Forest regression is an ensemble of unpruned regression trees created by bootstrap samples of the data with random feature selection in tree induction. It makes predictions by aggregating the predictions of the ensemble [37].
3. *Gradient Boosting Machine (GBM)*
   The Gradient Boosting Machine is an algorithm that generalizes the decision tree prediction model by allowing optimization of an arbitrary differentiable loss function [38, 39].

The common set of important attributes found using these models is then taken to be the set of critical technical attributes. The rationale behind taking the intersection of the important attributes is that the different approaches have different assumptions, weaknesses and strengths. Random Forest and GBM are decision tree based algorithms, which are robust to outliers in data points and deal well with irrelevant predictor attributes. ANN does not perform as well on the above characteristics, but is better at capturing non-linear and complex combinations of predictor attributes. For example, attributes A and B may not be important when taken alone, but may be significant when a combination of both is present. Random Forest and GBM may not consider A and B to be important attributes, but ANN will. Additionally, ANN and GBM may have some issues with over fitting, but Random Forests is more robust and will not over fit easily. All the algorithms picked can naturally handle both continuous and categorical predictor attributes, which is essential because the attribute list contains both binary and continuous data. They are also able to deal with incomplete data sets with some missing entries.

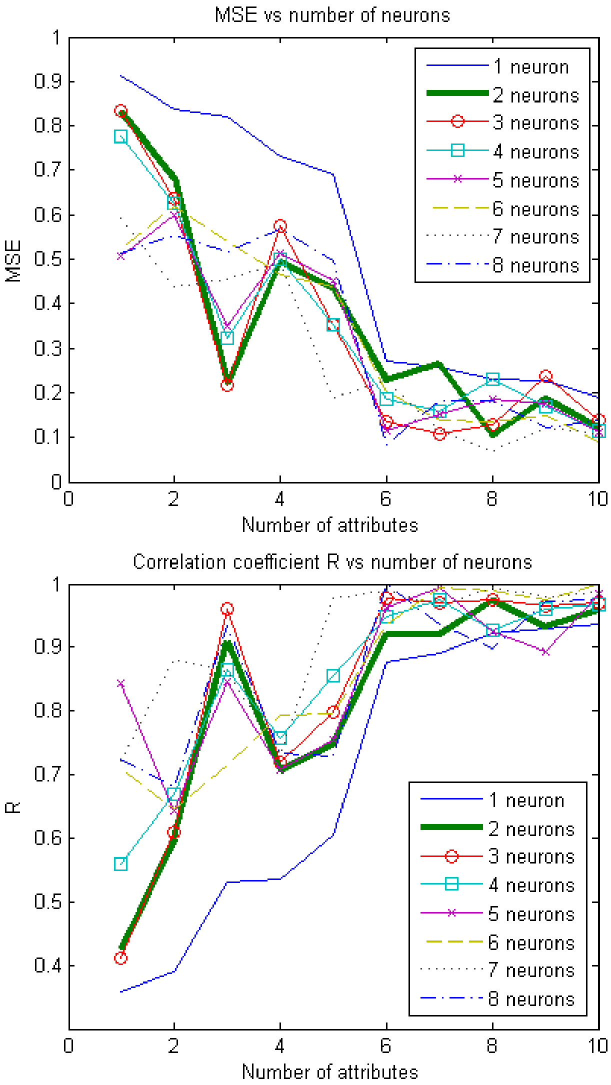

FIGURE 4. MSE AND R FITTING OVER 10 ATTRIBUTES USING 1-8 NEURONS

Artificial Neural Network Regression

A supervised feed forward Artificial Neural Network (ANN) fit was done in the MATLAB environment [40]. In order to determine the neural network architecture with an optimal number of neurons which gives the best fit without over fitting, the variance in the performance of fit with increasing neurons was tested. The number of neurons used for fitting was increased systematically from 1 to 8, using the top 10 attributes that mapped the best to market share. Each test was done with 300 trials to ensure that the global optimum was obtained, as MATLAB's neural network toolbox uses random initialization, which could affect the final result.

For each neuron number, the corresponding mean squared

TABLE 2. CORRELATION TABLE FOR IMPORTANT ATTRIBUTES FOUND BY ANN

R	1	2	3	4	5	6
Power Warranty	1.00					
Efficiency at STC	-0.38	1.00				
Time on Market	-0.19	-0.19	1.00			
NEC 2008	0.05	-0.04	-0.26	1.00		
ISO 9001	0.14	-0.02	-0.20	-0.12	1.00	
Weight per W	0.25	-0.86	0.17	0.00	0.00	1.00

error (MSE) and correlation coefficient R fits were obtained, and these were aggregated to form a graph of MSE and R fits using varying numbers of neurons as shown in Fig. 4. The optimal number of 2 neurons was selected, as it has a comparable MSE and R value to other neural networks with a higher number of neurons.

Using this optimal number of neurons for fitting, a new neural network model that maps each of the attributes to market share was built. Each optimization was run over 500 trials to ensure accurate optimization on the global minimum. MATLAB's parallel processing toolbox was used to run 3 processes simultaneously to speed up the analysis. The best model with the lowest MSE and highest corresponding R was picked to be the first element for the set of important attributes.

The second important attribute was chosen by creating new neural network models that map each attribute plus the first important attribute to market share. This was repeated until adding a new attribute did not reduce the MSE, resulting in a total of 6 important attributes. Further testing was conducted to ensure that the model is robust using correlation tables and bootstrapping methods. The corresponding bootstrapping values of MSE and R are displayed in Fig. 5. The correlation table of the important attributes is shown in Tab. 2.

Random Forest Regression

The Random Forest regression was performed using the `randomForest` statistical package created by Liaw and Wiener for the R Project environment based on the original Fortran code by Breiman and Cutler [41]. Since the Random Forest algorithm is robust to over fitting, very little tuning was required. The built in variable importance permutation calculation was used to identify critical attributes. 10,000 trees were grown and 3 variables were randomly sampled as candidates at each split. A lot of trees were necessary to get stable MSE and stable estimates of variable importance, as each input row needed to be predicted many times. The choice of 3 variables sampled at each split was decided by trying alternatives from 2-16 and choosing

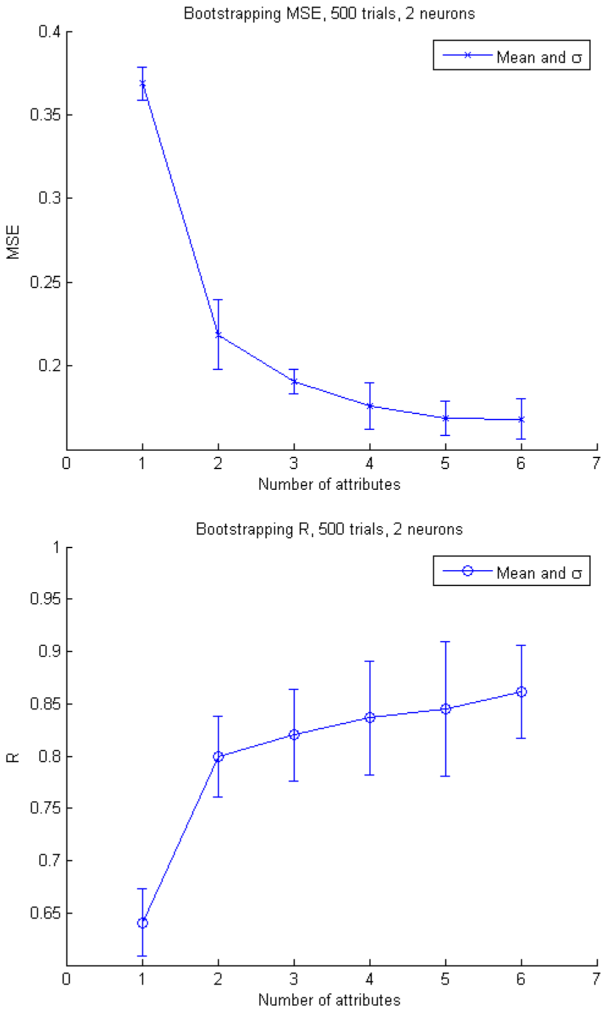

FIGURE 5. ANN BOOTSTRAPING ERROR VALIDATION

the best result with the lowest MSE.

100 regressions were done and an average of the importance values was taken, shown in Tab. 3. It was observed that although the variable importance measures varied slightly from run to run, the ranking of the importances was stable. Due to the nature of the method, cross validation was unnecessary as it generates an internal unbiased estimate of he generalization error as the forest building progresses. The importance threshold was chosen to be the absolute of the lowest importance value, resulting in a total of 13 important attributes.

Gradient Boosting Machine (GBM)

The gradient boosting machine was similarly performed in the R statistical environment using the `gbm` package written by Ridgeway based on extensions to Friedman's gradient boosting machine [42]. The learning rate, `shrinkage`, was set to 0.001,

TABLE 3. TOP ATTRIBUTES' RANDOM FOREST VARIABLE IMPORTANCE VALUES OVER 100 RUNS

Attributes	Mean	Std Dev
Time on market	19.72	0.78
Power variance (-%)	18.79	0.84
Weight per W	17.14	0.93
Power warranty	15.96	0.95
IEC 61215 (crystalline) or IEC 61646 (thin film)	11.28	0.80
Power ratio: PTC/STC	10.28	1.05
Safety class II @ 1000V	10.05	0.67
Efficiency at STC	9.62	1.05
Fill factor at STC	8.45	0.94
Power variance (+%)	7.77	0.91
Cell Number	7.20	1.00
IS0 9001	7.10	0.81
Workmanship Warranty	6.33	0.95

TABLE 4. TOP ATTRIBUTES' GBM RELATIVE INFLUENCE VALUES OVER 100 RUNS

Attributes	Mean	Std Dev
Weight per W	22.49	1.24
Power Variance (-)	18.74	1.06
Fill factor at STC	14.23	0.61
Efficiency at STC	12.21	1.00
Power ratio: PTC/STC	9.79	0.64
Effective Cost	6.29	0.56
Power warranty	2.28	0.22
Width	2.04	0.31

for the slowest rate but the highest accuracy. Other variables affecting the optimization, the maximum depth of variable interactions `interaction.depth`, the minimum number of observations in the trees' terminal nodes `n.minobsinnode` and the fraction of the training set observations randomly selected to propose the next tree in the expansion `bag.fraction`, were also varied systematically to obtain the optimum result with the lowest fitted MSE.

At each run, 4000 trees were grown with ten fold cross validation. The number of trees grown was chosen to be very high to be sure that the optimal number lies within the tested range. After each run, the function `gbm.perf` was used, which estimates the optimal number of trees using the data from the cross validation performed. The result at this number of trees is extracted and used.

The relative influence was then calculated by permuting one predictor variable at a time and computing the associated reduction in predictive performance. The computed relative influence was normalized to sum to 100. The mean of these relative influences over 100 regressions was then taken, shown in Tab. 4. The importance threshold was chosen to be a relative influence of 2, after which the relative influence values for the rest of the attributes holds steady around 1. This resulted in a total of 8 important attributes.

CRITICAL ATTRIBUTES

The summary of important attributes found from each method is shown in Tab. 5. The critical attributes are taken to be the important attributes that are common to all 3 methods, and form the feature set of concern. The rank ordering of the feature set is not considered to be important, as variations in the machine learning methods will cause differences in the rank ordering of the attributes.

The critical attributes found across all 3 methods are:

1. *Power warranty*
   Measure of power output performance guaranteed by the manufacturer over a period of time
2. *Efficiency at Standard Testing Conditions (STC)*
   Measure of performance of a panel
3. *Weight per W*
   Weight of panel per Watt of electricity produced, relates to ease of installation

At first glance, the critical attributes found are reasonable. Power warranty is linked to consumer confidence, as well as the reliability of the solar panel. Efficiency is a reflection of the performance of the technology, in this case the panel's ability to convert sunlight into electricity. Weight per Watt is a measure of the ease of installation of the panel, and hence points toward the influence of the installer on the purchase decision of the consumer.

It is important to note that the relationships between the critical attributes and market share derived from the machine learning algorithms do not imply causation. For example, the power warranty might not be the direct reason why customers prefer a certain panel over another, it might instead be a reflection of increased consumer confidence in the manufacturer's quality that results in increased market share. On the other hand, if there is no relationship, the attribute is not an important factor in the purchase decision.

TABLE 5. IMPORTANT ATTRIBUTES ACROSS 3 METHODS

Rank	ANN	RandomForest	GBM
1	**Power warranty**	Time on market (years)	**Weight per W**
2	**Efficiency at STC (%)**	Power variance (-%)	Power variance (-%)
3	Time on market (years)	**Weight per W**	Fill factor at STC
4	NEC 2008	**Power warranty**	**Efficiency at STC (%)**
5	IS0 9001	IEC 61215 / IEC 61646	Power ratio: PTC/STC
6	**Weight per W**	Power ratio: PTC/STC	Effective Cost/W ($)
7		Safety class II @ 1000V	**Power warranty**
8		**Efficiency at STC (%)**	Width (mm)
9		Fill factor at STC	
10		Power variance (+%)	
11		Cell Number	
12		IS0 9001	
13		Workmanship Warranty (years)	

Furthermore, the presence of all 3 critical attributes found does not guarantee market success for the product. The panel might have a good power warranty, high efficiency, and low weight per Watt, and still perform poorly on the market. Other non-technical factors like service quality, country-of-origin, and manufacturer reputation may play important roles in the purchase decision that are not reflected in this study. They will be taken into account in future work. What the analysis does show is that the panels need to have competitive levels of these critical attributes in order to have a chance at succeeding in the market. Hence, the list of critical attributes can be seen as ”must-have” attributes that designers should not neglect in the product development phase.

It is of value to note the factors that do not show up as important attributes in any of the methods. Interestingly, reduced waste in packaging, lead-free solder and the ISO 14001 environmental management standard fail to appear as important. The possibility that a consumer might miss these factors is low, because manufacturers heavily promote them as differentiating features, and they are displayed in large font at prominent places on the panel datasheets and advertising material. Since these are the only 3 factors in our analysis that reflect additional design thought on the eco-friendliness of the product, it can be inferred that consumers do not consider the environmental impact of non-core aspects of solar panels to be important when making their purchase decision. This is the opposite result of what is expected from using a stated preference method. This is a common problem in stated preference methods, with consumers responding differently in hypothetical situations than in actual market conditions. Homeowners who purchase PV systems frequently think of themselves as more environmentally conscious than the average population. However, previous research findings support our finding, showing that inconsistencies exist within ”green” consumption areas, where environmentally conscious consumers will not necessarily buy more ”green” energy products [43].

TABLE 6. R^2 VALUES FOR 3 METHODS

	ANN	RF	GBM
R^2	0.791	0.801	0.923

Unexpectedly, effective cost per Watt only appears in the GBM list of important attributes, although cost is frequently considered by many to highly influence the purchase decision. This result is a limitation of our study, as due to constraints in collecting data, we used the total cost of the PV system, which includes not only the panels, but also the inverter, labor, and installation costs, minus the state subsidy that was applied. This effective cost might not have been a factor of consideration when choosing between different panels. For a more accurate reflection of how cost influences this decision process, the panel price per Watt should have been used, but this data was unavailable in the California Solar Statistics database, and thus was not considered in this study.

Comparison of Methods

Some agreement between the various machine learning algorithms can be seen in Table 5. Only 3 attributes are common, 5 attributes occur twice, and 8 attributes only occur once. The different predictions are likely due to the noise in the data, which is an inherent problem when dealing with real data. The internal structure of the methods also differ, meaning the methods perform regression in differing ways. Although Random Forest and GBM are both decision tree based methods, because the learning approach differs, the important attributes found could be inconsistent. ANN has a completely distinct internal structure from the decision tree based methods, causing the important attributes found to be different. The combination of noisy real data and differing internal structures of the methods results in limited agreement.

A comparison of the accuracy of the models in predicting market share using the important attributes is shown in the R^2 goodness-of-fit values reflected in Tab. 6, where R^2 is calculated by Eq. 2.

$$R^2 = 1 - \frac{MSE}{\sigma_{MS}^2} \tag{2}$$

Table 6 indicates that all the models perform relatively well, with GBM being the most accurate. Ideally, Random Forest and GBM should have similar performance, because they are both decision tree based algorithms. The difference lies in how they optimize decision trees using ensemble approaches. Random Forest is usually more robust to internal parameter choice and performs well for wider varieties of parameters. Meanwhile, GBM tends to do better than Random Forest when the internal parameters are optimized carefully, as in this case. This highlights the need to carefully tune and test the parameters of machine learning methods before using the results.

With regard to computation time, GBM and Random Forest took a similar amount of time to run. ANN took a much longer time to train properly, although this might have been partly due to the difference in platform, with MATLAB running slower than R.

CONCLUSIONS

In this paper, we proposed a machine learning approach for revealing consumer preferences for technology products that are characterized by their technical attributes. We demonstrated this method with a case study on homeowner preferences regarding solar PV panels, and found 3 critical attributes that designers can prioritize for the optimization of time and resource allocation for the product development cycle.

Our main research questions can be answered in the following way:

1. *Can revealed consumer preferences be obtained from market data and engineering specifications using machine learning methods?*

   It appears that consumer preferences can be extracted successfully from marketing data and engineering specifications using the methods we attempted in this paper. However, as pointed out by prior work in the field, revealed preferences has the limitation in that only the set of attributes that are present in the data can be tested. There is a possibility that there are other critical attributes that are not present within this data set which are an important part of the homeowner purchasing decision process. Further work will include surveys as a data collection method for non-technical attributes, and will make use of stated preference methods to boost the existing revealed preference model [44–46].

2. *Is there agreement among the machine learning methods that suggest the validity of the data and methods?*

   There appears to be partial agreement among the methods, with noisy real data and differences in the internal structure of the methods causing this disparity. There is a need to look deeper into the machine learning methods we have explored in order to determine the differences in which the methods handle the data.

ACKNOWLEDGMENT

This work was supported in part by the Center for Scalable and Integrated Nanomanufacturing (SINAM) at UC Berkeley which is an NSF Nanoscale Science and Engineering Center. The opinions, findings, conclusion and recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Drucker, P., 1994. “The theory of the business”. *Harvard Business Review,* ***72***(5), Oct., pp. 95–104.

[2] Ulrich, K. T., 2011. *Product design and development*.

[3] Geroski, P., 2000. “Models of technology diffusion”. *Research Policy,* ***29***(4-5), Apr., pp. 603–625.

[4] Rogers, E., 1984. *Diffusion of innovations*. The Free Press, New York.

[5] Samuelson, P. A., 1938. “A note on the pure theory of consumer’s behaviour”. *Economica,* ***5***(17), Feb., pp. 61–71.

[6] Little, I. M. D., 1949. “A reformulation of the theory of consumer’s behaviour”. *Oxford Economic Papers,* ***1***(1), Jan., pp. 90–99.

[7] Samuelson, P. A., 1948. “Consumption theory in terms of revealed preference”. *Economica,* ***15***(60), Nov., pp. 243–253.

[8] Houthakker, H. S., 1950. “Revealed preference and the utility function”. *Economica,* ***17***(66), May, pp. 159–174.

[9] Szenberg, M., Ramrattan, L., and Gottesman, A. A., 2006. *Samuelsonian economics and the twenty-first century*. Oxford University Press, Nov.

[10] Mark, E., 1980. “The design, analysis and interpretation of repertory grids”. *International Journal of Man-Machine Studies,* ***13***(1), July, pp. 3–24.

[11] Tan, F. B., and Hunter, M. G., 2002. “The repertory grid

technique: A method for the study of cognition in information systems". *MIS Quarterly,* ***26***(1), Mar., pp. 39–57.

[12] Netzer, O., and Srinivasan, V., 2011. "Adaptive self-explication of multiattribute preferences". *Journal of Marketing Research,* ***48***(1), p. 140156.

[13] Marder, E., 1999. "The assumptions of choice modelling: Conjoint analysis and SUMM". *Canadian Journal of Marketing Research,* ***18***, pp. 3–14.

[14] Cohen, S., 2003. "Maximum difference scaling: Improved measures of importance and preference for segmentation". In Sawtooth Software Conference Proceedings, Sawtooth Software, Inc, Vol. 530, pp. 61–74.

[15] Green, P. E., Carroll, J. D., and Goldberg, S. M., 1981. "A general approach to product design optimization via conjoint analysis". *The Journal of Marketing,* ***45***(3), July, pp. 17–37.

[16] Green, P. E., and Srinivasan, V., 1990. "Conjoint analysis in marketing: New developments with implications for research and practice". *The Journal of Marketing,* ***54***(4), Oct., pp. 3–19.

[17] MacDonald, E. F., Gonzalez, R., and Papalambros, P. Y., 2009. "Preference inconsistency in multidisciplinary design decision making". *Journal of Mechanical Design,* ***131***(3), Mar., pp. 031009–13.

[18] Horsky, D., Nelson, P., and Posavac, S., 2004. "Stating preference for the ethereal but choosing the concrete: How the tangibility of attributes affects attribute weighting in value elicitation and choice". *Journal of Consumer Psychology,* ***14***(1 & 2), p. 132140.

[19] Cummings, R., Brookshire, D., Schulze, W., Bishop, R., and Arrow, K., 1986. *Valuing environmental goods: an assessment of the contingent valuation method*. Rowman & Allanheld Totowa, NJ.

[20] Kahneman, D., and Knetsch, J. L., 1992. "Valuing public goods: The purchase of moral satisfaction". *Journal of Environmental Economics and Management,* ***22***(1), Jan., pp. 57–70.

[21] Beshears, J., Choi, J. J., Laibson, D., and Madrian, B. C., 2008. "How are preferences revealed?". *Journal of Public Economics,* ***92***(89), Aug., pp. 1787–1794.

[22] Adamowicz, W., Louviere, J., and Williams, M., 1994. "Combining revealed and stated preference methods for valuing environmental amenities". *Journal of Environmental Economics and Management,* ***26***(3), May, pp. 271–292.

[23] Agard, B., and Kusiak, A., 2004. "Data-mining-based methodology for the design of product families". *International Journal of Production Research,* ***42***(15), pp. 2955–2969.

[24] Ferguson, C. J., Lees, B., MacArthur, E., and Irgens, C., 1998. "An application of data mining for product design". In IEE Colloquium on Knowledge Discovery and Data Mining (1998/434), IET, pp. 5/1–5/5.

[25] Kusiak, A., and Salustri, F., 2007. "Computational intelligence in product design engineering: Review and trends". *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* ***37***(5), Sept., pp. 766–778.

[26] Rucks, C. T., and Whalen, J. M., 1983. "Solar-energy users in arkansas: their identifying characteristics". *Public Utilities Forthnighly,* ***111***(9), Apr., pp. 36–38.

[27] Wander, J., 2006. "Stimulating the diffusion of photovoltaic systems: A behavioural perspective". *Energy Policy,* ***34***(14), Sept., pp. 1935–1943.

[28] Faiers, A., and Neame, C., 2006. "Consumer attitudes towards domestic solar power systems". *Energy Policy,* ***34***(14), Sept., pp. 1797–1806.

[29] Jetter, A., and Schweinfort, W., 2011. "Building scenarios with fuzzy cognitive maps: An exploratory study of solar energy". *Futures,* ***43***(1), Feb., pp. 52–66.

[30] Griffin, A., and Page, A. L., 1993. "An interim report on measuring product development success and failure". *Journal of Product Innovation Management,* ***10***(4), Sept., pp. 291–308.

[31] Griffin, A., and Page, A. L., 1996. "PDMA success measurement project: Recommended measures for product development success and failure". *Journal of Product Innovation Management,* ***13***(6), Nov., pp. 478–496.

[32] CSI, 2011. California solar initative: Current CSI data. http://www.californiasolarstatistics.org/current_data_files/, Sept.

[33] SEIA, 2011. U.S. solar market insight. Executive summary, SEIA/GTM Research.

[34] Cortes, C., and Vapnik, V., 1995. "Support-vector networks". *Machine Learning,* ***20***(3), Sept., pp. 273–297.

[35] Obrien, R. M., 2007. "A caution regarding rules of thumb for variance inflation factors". *Quality & Quantity,* ***41***(5), Mar., pp. 673–690.

[36] Rojas, R., 1996. *Neural networks: a systematic introduction*. Springer.

[37] Breiman, L., 2001. "Random forests". *Machine Learning,* ***45***(1), Oct., pp. 5–32.

[38] Friedman, J., 2001. "Greedy function approximation: a gradient boosting machine". *Annals of Statistics*, p. 11891232.

[39] Friedman, J., 2002. "Stochastic gradient boosting". *Computational Statistics & Data Analysis,* ***38***(4), p. 367378.

[40] Beale, M., and Demuth, H., 1998. "Neural network toolbox". *For Use with MATLAB, Users Guide, The MathWorks, Natick*.

[41] Liaw, A., and Wiener, M., 2002. "Classification and regression by randomForest". *Resampling Methods in R: The boot Package,* ***2***(3), Dec., pp. 18–22.

[42] Ridgeway, G., 2007. "Generalized boosted models: A guide to the gbm package". *Update,* ***1***, p. 1.

[43] Laroche, M., Bergeron, J., and Barbaro-Forleo, G., 2001.

"Targeting consumers who are willing to pay more for environmentally friendly products". *Journal of Consumer Marketing,* ***18***(6), Jan., pp. 503–520.

[44] Dietrich, E., 2002. "Combining revealed and stated data to examine housing decisions using discrete choice analysis". *Journal of Urban Economics,* ***51***(1), Jan., pp. 143–169.

[45] Brownstone, D., Bunch, D., and Train, K., 2000. "Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles". *Transportation Research Part B: Methodological,* ***34***(5), p. 315338.

[46] Hensher, D., and Bradley, M., 1993. "Using stated response choice data to enrich revealed preference discrete choice models". *Marketing Letters,* ***4***(2), p. 139151.