8-2018

# A Cluster Based Model to Enhance Acceptance of New Energy Driven Technologies

Mohammad Ali Asudegi
*University of Tennessee*, masudegi@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Mohammad Ali Asudegi entitled "A Cluster Based Model to Enhance Acceptance of New Energy Driven Technologies." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial Engineering.

Rapinder S. Sawhney, Major Professor

We have read this dissertation and recommend its acceptance:

David L. Greene, John E. Kobza, Andrew Yu

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

# A Cluster Based Model to Enhance Acceptance of New Energy Driven Technologies

**A Dissertation Presented for the**

**Doctor of Philosophy**

**Degree**

**The University of Tennessee, Knoxville**

**Mohammad Ali Asudegi**

**August 2018**

# Abstract

Resistance against new innovative technologies by customers has been studied in many publications to improve prediction of behavior. Econometrics models, the Technology Acceptance Model by Fred D. Davis (1989), and market research models are the most widely used modeling techniques to predict and understand customer behaviors. The proposed methodology in this paper advances current models by relaxing many of their assumptions and increasing prediction accuracy. A case study in predicting hybrid car buyer behaviors is performed to illustrate and validate the suggested modeling method named as the Energy Efficiency Technology Acceptance Model.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1: Introduction

Concerns about global warming and increase in the price of energy are the main reasons for researchers to study different innovative solutions to increase the efficiency of energy driven industries and machines. In 1987, the United Nations World Commission on Environment and Development (WCED) defined sustainability as “development that meets the needs of the present without compromising the ability of future generations to meet their own needs.” Those who wish to increase sustainability look to conserve resources and the environment for future generations by advancing current technologies or introducing new innovative products that deplete less energy and emit fewer harmful substances. Acceptance of these innovations by customers is as important as their introduction to the market to decrease production of greenhouse gases and improve sustainability. Unfortunately, resistance to innovation is consumers’ reaction to new or improved technologies and products that come into the market. According to C. Merle Crawford (2008), 90% of new products do not survive on the market. Increasing the success probability of innovative products needs better communication of new technologies to the market and improved focus of resources on the right customers. This requires prediction of who will accept new technology and a better understanding of the motivations of different categories of buyers. This study introduces a novel model to predict the acceptance of new innovative technologies reducing energy consumption. This can help manufacturers and policy makers in the field of sustainable energy to improve market share of new more efficient technologies. The proposed model is applied in sustainable transportation for evaluation.

## 1.1. Problem Statement

Many models have been developed to understand customer choice and motivations in order to predict customers of new products. Developed models are mostly econometrics

models and assume customers to be informed Economy Rational decision-makers who think and behave similarly (Bento, Li, and Roth, 2012). An informed Economy Rational customer is defined in current studies as an individual who has enough knowledge about goods and who performs calculations to evaluate choices. This customer will choose goods which benefit him the most instead of choosing another option. The amount of benefit received is calculated by a utility function. A utility function measures the monetary value of gain to the cost of choices. The present value of a future monetary gain is always lower than the gain itself. This is a fact in financial management and is critical to calculating the rate of return on loans with a perspective of the difference in the value of money in the present and in the future. This means a gain will be more valuable if received in a shorter period. Researchers used this concept to understand the gap between acceptance of innovative efficiency technologies in the real word and predicted acceptance by econometric models. They assume that customers look for a discount in future saving of energy if they need to pay a price premium for an innovative energy efficiency technology (Hirst, 1990). Based on the assumption of the informed Economy Rational customer, many studies calculated the implied discount rate and payback periods to understand and predict the market. While many concluded there is a high implied discount rate of return by customers, other studies resulted in low implied discount rate of return (Wolverton, 2011; Gallego et al., 2013). Many theories have been used to understand this outcome including Energy Paradox and Loss Aversion, but no study could make an end to this source of conflict. High implied discount rate of return can be explained by the theory of Energy Paradox and Loss Aversion. Low implied discount rate of return can be explained by the tendency of individuals to be risk adverse (Bento, A. M., et al., 2012). While econometric modeling is the most well-known technique to predict acceptance of new products, many researchers, including Kahneman (2011), question its validity and the assumption of Economy Rational customers (Greene, 2014). Even the widely used McFadden Discrete Choice Model works only under bounded conditions. In the real world, most customers do not have enough information about new products and do not perform complicated mathematical calculations for choosing products. Instead, they use heuristic easy decision-making methods (Kahneman, 2011).

Turrentine and Kurani (2007) showed that customers make decisions based on their impressions and feelings. Indeed, Davis (1989) had introduced earlier a non-econometric model known as the Technology Acceptance Model (TAM) based on this assumption. This model was primarily developed to understand resistance of users and customers in the field of information technology. According to TAM, perceived usefulness will result in accepting new technologies. While TAM omits the assumption of fully informed rational individuals, it only explains a small portion of new technology acceptance. The weak prediction power of the model has been mentioned by many researchers including Legris, Ingham, and Collerette (2003). According to behavioral specialists, humans who are fully informed may make biased decisions (Andrew J. Barne, 2016). This is key to why predicting acceptance of new technologies is a challenge. Humans are also biased differently because of environmental factors, and this affects their decision processes. While differences in individual and environmental attributes indicate the possibility of different decision processes for individuals, current studies, including Davis's TAM (1989), do not properly address the heterogeneity of consumer decision-making. Also, the limited number of analyzed factors have been inadequate to overcome the complicated behavior of customers. Indeed, none of the existing models are even capable of comparing the importance of diverse attributes which have already been analyzed.

In addition to economists, manufacturers and retailers are also interested in predicting acceptance of new technologies. In the field of market research, data is captured through surveys and designed experiments or is extracted from alternative available sources to provide market related answers like who and where the customers are. Developing a hypothesis is an important part of market research. A hypothesis is tested using statistical and data mining tools. Researchers may also use econometric models such as Discrete Choice by McFadden or non-econometric models such as Technology Acceptance by Davis. These models are divided into two categories based on their source of data. One category includes models which use existing data from other studies, and the other category consists of models which capture their own data through designed experiments or surveys. Capturing data for analysis is a well-known challenge in studying acceptance of new technologies. Buying a new innovative product is considered a rare event. This makes the process of data collection and analysis more difficult. The process of capturing

data for a rare event is frustrating and expensive and needs to be addressed with a new systematic data collection method.

## 1.2. Approach

In this study, an innovative prediction model for acceptance of new energy efficiency technology is introduced using a new perspective of the problem. Many previous assumptions are eliminated by clustering of the customers. Customers are not assumed to be informed Economy Rational individuals. They are not assumed to make mathematical calculations to choose the good which maximizes their utility function. Also, the model considers heterogeneity of individuals and their decision processes. The proposed technique is the first to use a non-parametric probabilistic model to predict acceptance of new innovative products and simulate human decision processes. The model not only predicts the probability of acceptance for different clusters of customers, but also highlights their motivations and decision processes. The resultant model is a more robust and reliable prediction model compared to current ones with respect to understanding market opportunities and customers' motivations and their preferred channels of communication to improve the market share of innovative energy efficiency technologies.

Figure 1.1 shows the approach to address and solve the stated problems in this research and introduces the new model. The introduced model in this study is referred to as the Energy Efficiency Technology Acceptance Model (ETAM).

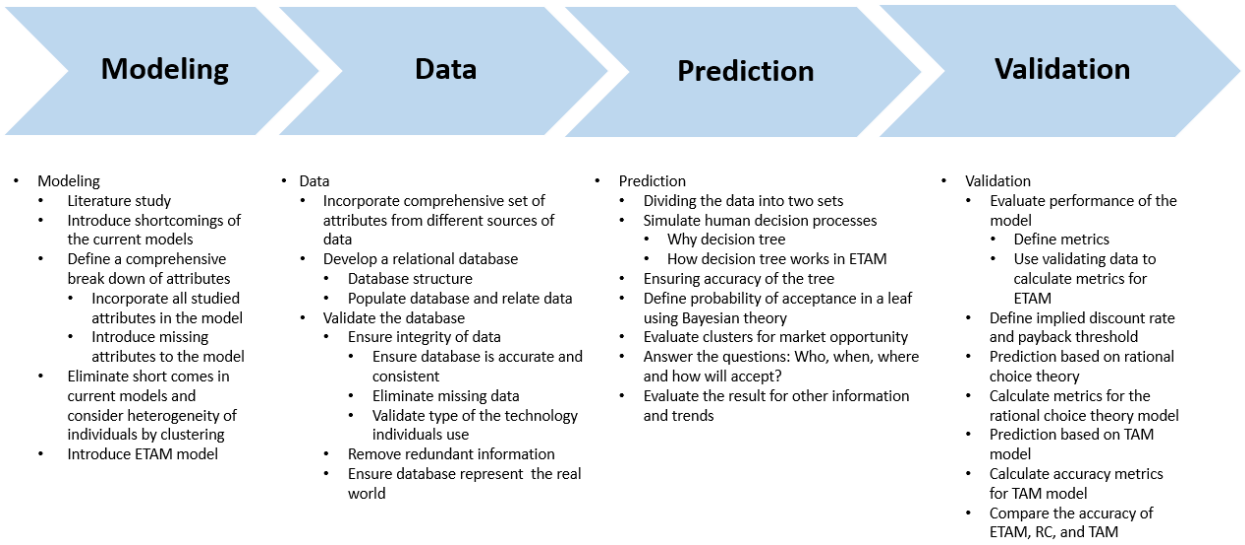



*Figure 1.1: Approach*

In the process of developing ETAM, empirical models are evaluated to discover their assumptions, shortcomings, and studied factors since 1900. Input attributes of the new model are identified to address the limited number of attributes in current models. Analysis of the input attributes helps to assure their importance. These attributes not only include what customers indicate and think and believe, but also their actual behavior. This uncovers both stated and revealed preference of customers. In addition, importance comparison of the input attributes is possible through the comprehensive breakdown of attributes and their use for prediction of technology acceptance by individuals.

The data collection and analysis part of the proposed model, ETAM, is able to handle a high number of input attributes without the need to reduce dimensionality of data.

The outcome of the model is validated, and evaluation metrics are defined. It is important that the introduced metrics be applicable to both empirical and other models for comparison. ETAM suggests using metrics based on confusion matrix via the small group of data which has not been used in the process of developing the prediction model. In addition, a case study is done to compare the result of the model with current empirical models of choice.

## 1.3. Methodology

A comprehensive set of attributes from various online data sources are captured to describe the categories of inputs introduced in the previous section. Captured data consist of individuals' demographic information, their purchase history, their environment, their behavior, their use of the product, their beliefs, and their viewpoints. Additionally, available federal and local laws or incentives to buy the innovative technology are considered.

A data structure and relational database are developed to store and relate captured data from various sources. This is preferable considering the size of data in this effort. Data is validated and checked for accuracy before use as input to the model. Instead of deleting incomplete records, the introduced model uses them to investigate the possibility of existing meaningful trends in missing fields of data. Data is aggregated, matched, and cleaned as necessary.

ETAM uses publicly available data which reduces the difficulties with collection. A considerable amount of data is available online, and it can be used by researchers for a low cost or free of charge. While using available online data reduces the cost of analysis, there are two major problems with this data. Acceptance of an innovative technology is a rare event. A rare event means that we have very few recorded observations of acceptance of the innovative technology by individuals. In addition to this, it is possible that the available data does not represent the real world. These problems are addressed in models by applying weights to records.

For model training and validating, data is divided into two groups by the ratio of 4:1. The bigger group of data is used to make clusters of customers using the supervised decision tree clustering technique. The motivations, assumptions and characteristics of each cluster of customers are highlighted, and the probability of acceptance of the innovative technology by individuals in each cluster is calculated using Bayesian Theory. Clusters will be evaluated as probable market opportunities. The smaller group of data is used for validation.

Previously used statistical methods could not handle a high number of input attributes, especially with multicollinearity. This resulted in a limited number of attributes and an inability to measure their relative importance. To solve this problem, and to address

heterogeneity of consumer decision-making processes, this study uses the decision tree technique to cluster the customers using the input factors of the model. Decision tree mimics the human process of thinking and decision making. The decision tree analysis used here divides entities, which are individuals in this research, into groups in such a way that individuals in each group are more similar and have the same probability of accepting the new technology. This idea is illustrated in Figure 1.2, which shows a sample of people divided into four clusters. Input attributes define the border between clusters. A pure cluster consists of individuals who all accepted or rejected the innovative technology, and Impurity is a measure of how different a cluster is from a pure cluster. The objective function of the decision tree minimizes the sum of impurity of response in nodes of the tree, expressed by Equation 1.1.

$$\text{Min} \sum \mathrm{I_j} \tag{1.1}$$

$\mathrm{I_j}$ Impurity in node *j*

Individuals in each node of the output tree have similar characteristics, motivations, and behaviors. An ideal tree places all individuals who accepted the innovative technology in one cluster and the others in another one. In this ideal situation, all those who accepted the new technology have the same characteristics and motivations. Details of how impurity is calculated and how the algorithm works can be found in Chapter 4. Impurity is measured using the response variable. Having more individuals who accepted the technology in a node results in a higher impurity of this response.

Because each customer has different assumptions and levels of risk aversion (Rogers, 1962) and can be biased differently when making decisions, the clustering technique is inevitable to identify groups of customers and their motivations. This removes boundaries in previous models and makes the suggested model more robust and reliable than empirical ones when applied to predict acceptance of new products.


*Figure 1.2: Customer Clusters*

ETAM would help to increase the probability of accepting the new technologies among early adopters by highlighting the effective communication channels and motivations for the right category of customers. This predictive model can help policy makers and manufacturers to have more reliable in-depth knowledge about customers and their choices.

## 1.4. Outline

Chapter 2 covers a review of literature, empirical models, their assumptions, and studied attributes. It will discuss current models in addition to a comprehensive comparison of empirical models with regard to their input attributes and prediction techniques.

Chapter 3 consists of three main parts. The first part of Chapter 3 will introduce an innovative comprehensive breakdown of attributes in predicting acceptance of new technologies, which will be used as the guideline for selecting input attributes for ETAM. This breakdown includes categories of previously studied attributes and new ones. In the second part of Chapter 3, ETAM will be introduced. Details of data collection and the use of data to predict in ETAM will also be discussed in the second part of this chapter. In the third and last part of Chapter 3, evaluation metrics are defined for ETAM. In addition, two

more empirical models will be introduced. These can be used for performance comparison of ETAM.

Chapter 4 includes a case study and implementation of ETAM in a real word problem. Online available data are captured and used as the input of the model to predict acceptance of hybrid cars and their market opportunities.

Chapter 5 evaluates the results of this study including ETAM and the case study.

# Chapter 2: Literature Review

In this chapter, the literature study regarding empirical econometric models, the revolution of econometrics models, Energy Efficiency Gap, and modeling in market research will be presented. Researchers in the fields of both economics and marketing have tried to understand customer behavior and predict acceptance of products, but none could prove a reliable technique (Bento, Li, and Roth, 2012). Econometric models are based on the Rational Choice Theory. When econometric models were applied to predict energy efficiency choices, researchers found a big gap between the predicted acceptance rate and the real world acceptance rate. They tried to reason this variation by introducing Energy Efficiency Gap (Hirst, 1990). In market research, researchers not only use statistical techniques, but may also use Discrete Choice Model from economics. Another well-known non-econometric model to predict acceptance of new technologies is TAM by Davis (1989). Figure 2.1 shows a summary of studied models, including considered attributes, and their differences, which will be discussed in detail in this chapter. In the following sections, a review of the models will be presented as the basis to develop ETAM. At the end, a review and summary of studied attributes in empirical studies will be provided. These will be used later as probable significant factors in acceptance of new efficiency technologies.

## 2.1. Revolution of Econometric Modeling

Many models have been introduced for predicting behavior of customers. Predicting customers of commodities has always been an attractive topic to economists. In addition, a well-known problem is rejection of new technologies or systems which require users to perform a specific job in a new and different way.

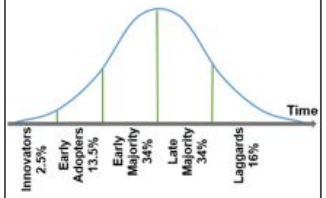
| | Models in ECONOMIC RESEARCH for new product acceptance | | | Models in MARKET RESEARCH for new product acceptance | | Studies for ENERGY EFFICIENCY choices |
|---|---|---|---|---|---|---|
| | **Diffusion of Innovations by Rogers 1962** | **New Theory of Consumer Demand by Lancaster 1966** | **Random Utility Theory, Discrete Choice Modelling by McFadden 1976-1999** | **TAM by Fred D. Davis 1989** | **Overcoming Barriers To Energy Conservation by Blumstein et al. 1980** | **Energy Efficiency Gap** |
| **Description** | Diffusion of Innovation is the process by which an innovation is communicated over time among the participants in a social system. | What consumers are seeking to acquire is not goods themselves but the characteristics they contain | Model the decision process of an individual via revealed preferences or stated preferences (A over B; B over A, B & C) | Perceived usefulness result in tendency to accept the new technologies | Introduced barriors to energy conservation | Customers undervalue the future fuel savings (Hirst 1990) Implied rate by customers is as high as 25% (Hausman 1979) |
| **Model/Theory/Research** |   | $\text{Max } z_j = \sum_k b_k x_{jk}$ <br> $b_k$ Scaler or importance of characteristic k <br> $x_{jk}$ Amount of characteristic k in product j <br> $z_j$ Amount of gain from product j | $z_j = \sum_k b_k x_{jk} + \varepsilon_j$ <br> $\varepsilon_j$ New introduced error term <br> $\hat{P}_j = \frac{e^{z_j}}{\sum_I e^{z_j}}$ <br> $\hat{P}_j$ Choice probability of good j | $P = \frac{e^{a+\beta v}}{1+e^{a+\beta v}}$ <br> P Probability of acceptance <br> β Vector of scaler <br> a Parameter of the regression <br> v Vector of perceived views | Market study through survey | $pv = \sum_0^n \frac{Y}{(1+r)^n}$ <br> n Length of payback <br> pv Present value <br> r Implied discount rate <br> Y Future saving in energy |
| **Type** | Theory | Deterministic, parametric model | Probabilistic, parametric model, analysis of preferences needs more data than observable market data | Probabilistic, non-parametric model | Study | Deterministic, parametric model |
| **Assumptions** | Acceptance is the result of communication | Informed Economy Rational customer | Informed Economy Rational customer | Acceptance is the result of perceived view | Rejection is the result of social and institutional barriers | Principal of Loss Aversion theory, customers expect discount |
| **Rational Choice Theory Consideration** | No | Yes | Yes | No | Yes | Yes |

*Figure 2.1: Comparing Empirical Models*

All econometric models are based on the Rational Choice Theory (Savage, 1954). The Classical Theory of Customer Demand assumes that customers try to maximize their utility or profit by choosing a given product over its alternatives. This assumption is valid only if the customers are Economy Rational; that is, they have enough knowledge about the goods and calculate their cost and profit (Simpson, 1974). Classical econometric models are deterministic and assume that all individual have enough knowledge regarding the product and calculate the cost of ownership. This is known as the informed customers assumption. This assumption says that a customer chooses the good which maximizes his or her utility or profit. The majority of economists try to formulate the behavior of individuals using parametric methods. Equation 2.1 shows the objective function of this classical type of modeling. The model will pick the product which maximize the net profit which is the difference between the gain and cost.

$$\text{Max } U(z): \quad z_j = g_j - c_j \tag{2.1}$$

$c_j$ Cost to own product *j*

$g_j$ Total monetary gain or total profit from product *j*

$z_j$ Net profit from product *j*

The model chooses product *j* which gives the highest net profit for the individual.

Rogers (1962) introduced the Diffusion of Innovations Theory. This theory shows how new technologies spread by dividing individuals into innovators, early adopters, early majority, late majority, and laggards. Based on this theory, social status, geography, education, and information are attributes affecting the spread of new technologies. While this theory was a revolution in economics and introduced a new perspective of acceptance rate in different intervals, it was not put into the majority of econometric models because they were deterministic. Figure 2.2 shows the expected amount of market share acquired by each of five categories of customers based on the Diffusion of Innovations Theory.

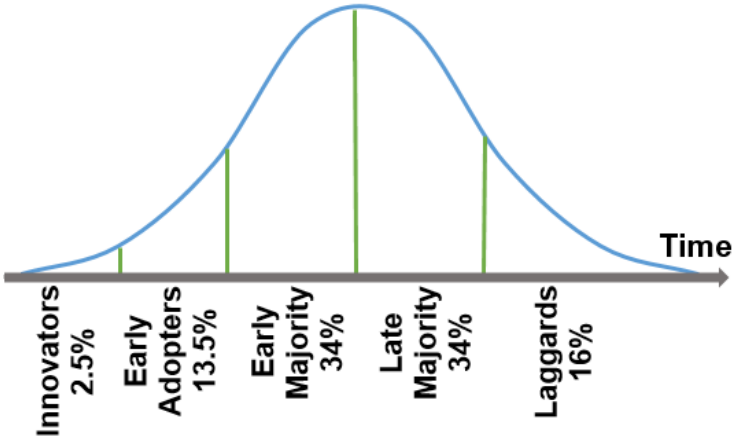



*Figure 2.2: Diffusion of Innovations Theory*

Innovators are more risk seeking than average and are willing to be first adopters of new technologies. According to Rogers' theory, these individuals are the first 2.5% of adopters.

The second group of individuals are early adopters. Individuals in this group have a high level of leadership. These individuals have higher social status, education, and income. Early adopters are expected by the Diffusion of Innovations Theory to be a total of 13.5% of all the market share.

The third group of individuals is early majority. This group of customers is believed to have higher than average social status and contact with others who have already accepted the technology. Acceptance of the new technology by this group brings the market share of the product to 50% of the total market share.

The fourth group of adopters is named late majority. They adopt the new technology only when the majority have already accepted it. They are believed to have lower than average social status, education, and income. They are conservative and do not trust new technologies. The size of this group is believed by the theory to be 34% of the market share.

The fifth group of adopters is called laggards. These individuals have the highest amount of resistance to change. They are believed to have the lowest social status, income, and highest age by the theory. The size of this group is 16% and by their accepting innovation, the market share reaches its full amount of 100%.

There are many factors that support Diffusion Theory. Manufacturers learn how to improve products over time. They may reduce the price and increase the quality of the same product. Their customers will see less risk in accepting a product which has been accepted by their friends or those who are in contact with them.

Kelvin Lancaster introduced the New Theory of Customer Demands in 1966 (Hendler, 1975). The New Theory of Customer Demand by Lancaster indicates that consumers are looking to receive the characteristics of the goods. While the model is still a deterministic parametric one, it is the first model that considers the characteristics of the product as important factors in acceptance. Similarly to previous studies, this model assumes that customers are informed Economy Rational individuals and that they know the characteristics of the available products and technologies.

Equation 2.2 shows the objective function that customers try to maximize by making rational choices according the New Theory of Customer Demand. The product with the characteristics that maximize the value of the function is the predicted candidate to be chosen by the customer.

$$\text{Max } U(z): \quad z_j = \sum_k b_k x_{jk} \tag{2.2}$$

$b_k$ Scaler or importance of characteristic k

$x_{jk}$ Amount of characteristic k in product j.

$z_j$ Amount of gain from product j.

The scaler is unknown and needs to be calculated for the model. Not all characteristics have the same value for the customer. The scaler in the model is used to adjust the importance of the characteristics in the eyes of the customer. The product which will give the highest relative total of gain from its characteristics is chosen by the model as the predicted decision of the customer.

Both the Classical and the New Theory of Consumer Demand use parametric and deterministic models to understand customer behaviors and choices. McFadden (1976) introduced the Random Utility Theory. He added a randomness term to the Lancaster

model which resulted in a probabilistic parametric model. Discrete Choice Modeling, which is derived from the Random Utility Theory, is widely used to measure preference of customers. This is the first probabilistic econometrics model to predict acceptance by customers.

The Discrete Choice Model is used to understand revealed or stated preferences of customers regarding characteristics of a product. For example, in a survey, the researcher would ask customers to rate product A over B in one question and in the next question ask the customer to rate product B over product A, B, and C. As the result, the model can position products A, B, and C for the customer. The products A, B, and C have different characteristics, and the goal is to understand which characteristics are more important in choosing products. Analysis of preferences needs more data than observed data from the market. In addition to economy studies, this model is widely used in market research by manufacturers to improve the design of their products and expand the market share of their products.

This model assumes that customers are aware of product information and systematically weigh their characteristics. While the model still assumes that customers try to maximize their utility function, it improves previous models by being probabilistic and adding an error term to the model.

Equation 2.3 shows the McFadden Random Utility Theory which is the updated version of the utility function in the New Theory of Consumer Demand discussed in the previous equation.

$$\text{Max } U(z): \quad z_j = \sum_k b_k x_{jk} + \varepsilon_j \tag{2.3}$$

$b_k$ Scaler or importance of characteristic k

$x_{jk}$ Amount of characteristic k in product j

$z_j$ Amount of gain from product j

$\varepsilon_j$ New introduced error term

If an individual selects choice $a$ over other alternatives, it means that the utility of this choice, $z_a$, is greater than the amount of utility from other choices. This helps to calculate the unknown scaler $b_k$, which shows the importance of the characteristic $k$ for the individual, and term $\varepsilon_j$ as the error in choosing or not choosing product $j$ by the individual. If the model fits perfectly and all chosen goods are predicted perfectly, then the error term is equal to 0, and the model will be the same as the model in the New Theory of Consumer Demand.

McFadden received the Nobel Memorial Prize in Economics for the development of the Random Utility Theory and the Method for Discrete Choice Analysis. Equation 2.4 shows how the probability of acceptance is calculated in the Discrete Choice Model.

$$\widehat{P}_j = \frac{e^{z_j}}{\sum_J e^{z_j}} \tag{2.4}$$

$\widehat{P}_j$ Choice probability of good $j$

$z_j$ Amount of gain from product $j$

Choice probability, $\widehat{P}_j$, is the probability of an individual choosing good $j$ over other goods. The numerator is e raised to the power of the utility for good $j$ and the denominator is the sum of e raised to the power of utility for all available options. While many studies have been done in the field of economics to improve the prediction power of the models on the basis of the Theory of Consumer Demand, there exits another category of studies which are on the basis of TAM. TAM has not been accepted in economics but is used widely in market research.

## 2.2. TAM

Fred D. Davis (1989) suggested that the perceived usefulness of a new technology by customers results in a tendency to accept or reject it. Individuals have views consisting of perceived usefulness, cons, and pros of new technologies. Davis's model is known as

TAM. TAM uses perceived attributes as input. Davis also indicated the effect of external factors on individuals' perceptions.

TAM mainly has been developed for acceptance of new information technologies. While this model does not assume the Economy Rational thinking process of customers, it still has very limited prediction power. Legris, Ingham, and Collerette, (2003) did a comprehensive study and concluded that TAM can only explain 40% of technology acceptance.

Combining the Theory of Consumer Demand and TAM results in a new utility function similar to Lancaster's model in Equation 2.2. The only difference is the input attributes to the model. The independent attributes used as the input of the Lancaster Utility Theory model. $x_{jk}$ should be replace by the perceived view-points of the customers.

## 2.3. Energy Efficiency Gap

While the development of econometric models helps policy makers to understand the market, their accuracy and reliability have not been proven in predicting acceptance of innovative efficiency technologies. Many researchers, including Jaffe et al. (1994), say this gap exists because customers undervalue the future savings they will receive. The efficiency gap is illustrated by a comparison of the market interest rate and implied discount rates by customers to choose energy efficiency technologies (Hausman, 1979). According to the Utility Theory, consumers experience satisfaction from goods, but it is not possible to measure this satisfaction directly. The utility function used in econometric Rational Choice models monitors the monetary value of different choices or their characteristics for customers. The present value of a future monetary gain is always lower than the gain itself. This is a fact in financial management and is critical to calculating the interest on loans with a perspective of the difference in the value of money in the present and in the future. The same concept exists in econometric modeling. Many researchers studied the amount of premium price customers are willing to pay for owning an innovative energy saving technology (Hausman,1979; Wolverton, 2011; Gallego et al., 2013). This is essential for predicting the acceptance and market share of an innovative technology that should compete with existing technologies. Customers consider the money saving

from the more efficient technology as a gain in the utility function and the premium cost to own it as a loss. The application of the present value of future saving of money to calculate willingness to pay a premium for a more efficient technology by customers can be seen in Equation 2.5. The main difference between Equation 2.5 and Equation 2.1 is that the value of gains decreases as time passes. This means that gains will be more valuable when received sooner and later gains will be less valuable even if they are equal.

$$z_j = pv_j - c_j = \sum_{0}^{n} \frac{y_j}{(1+r)^n} - c_j = \sum_{0}^{n} \frac{u \times d_j \times e}{(1+r)^n} - c_j \quad (2.5)$$

$c_j$ Cost to own product j

$d_j$ Amount of decrease in the unit of energy consumption from choosing product j

$e$ Unit cost of energy

$n$ Product service life time

$pv_j$ Present value of future saving in energy from choosing product j

$r$ Implied discount rate

$y_j$ Future saving in energy from choosing product j

$z_j$ Amount of premium a customer is willing to pay for product j

The implied discount rate, $r$, is the same as the rate in calculating the present value of money which will be received in the future. This rate is defined by customers and is their expectation from their investment. This unknown rate needs to be calculated. $z_j$ is the amount of money an individual is willing to pay for the more efficient technology. The money which will be received in the future from technology $j$ is indeed the amount of money being saved as the result of choosing the more efficient technology. This value is calculated by multiplying the yearly amount of technology usage, the amount of decrease in the unit of energy consumption from choosing the efficient technology and the unit cost of energy.

The implied discount rate by customers has been studied in many publications, and many researchers, including Helfand, Wolverton (2011) and Greene (2011), showed a high

implied discount rate by customers. Others, including Francisco Gallego et al. (2013), showed low implied discount rate by customers.

Researchers use different psychological theories for supporting the low or high implied discount rate including Energy Paradox and Loss Aversion (Bento, A. M., et al., 2012). The Energy Paradox Theory indicates that most customers will undervalue the future savings of more efficient technology. Energy Paradox Theory can exist due to customers' lack of information and mathematical skills. The Economic Principal of Loss Aversion Theory refers to the tendency of humans to overweigh the loss over the gain. For example, a person might weigh a 10% probability of losing $100 as more significant than a 10% probability of gaining $100. Loss aversion is one of the main motivations for people to buy insurance higher than the expected cost due to loss. This theory can be used to answer why many customers, even fully informed ones, prefer not to pay the upfront higher cost of a more efficient option in favor of future gain. Uncertainty of future energy price may be another reason individuals are unwilling to pay for the premium of a new more efficient technology. A risk adverse individual is willing to pay, potentially losing currently available money, for assurance against future loss, while a risk seeking individual is willing to invest in hope of future gain. Since individuals have different levels of risk seeking tendency, Loss Aversion results in different behavior and implied discount rate among them.

Higher risk aversion results in higher implied discount rate (Lam Weng Siew et al., 2014). While Rogers's Diffusion of Innovations Theory shows a relation between implied discount rate as an indicator of risk tolerance and social status, income, geographic location, and education, Klapper et al. (2005) and Greene (2011) showed that there is no correlation between Loss Aversion and any single selection or group of social, economic, or demographic attributes.

While Gilbert E. Metcalf and Kevin A. Hassett (1993) concluded that implied discount rate by customers is the result of rational thinking, Kahneman (2011), Turrentine, and Kurani (2007) showed that most customers do not use a rational calculating process to choose a product. Howarth and Sanstad (1995) reviewed the econometric models and concluded that they fail to calculate the discount rate.

## 2.4. Market Research

In addition to economists, manufacturers and retailers are also interested in predicting acceptance of new technologies to improve their businesses. Market research looks to answer where and who the customers of a product are and when they are probably going to buy the product. The researcher defines a problem or questions which need to be answered. This is followed by the researcher developing a hypothesis which needs to be tested to answer the questions. Then, the process of data collection is designed and evaluated. The best data analysis tools are selected to evaluate the data and accept or reject the hypothesis. The selected hypothesis is usually tested by statistical tools such as the t-test, z-test, and f-test. Econometric models such as Discrete Choice by McFadden (1976) or TAM by Davis (1989) are also widely used to answer questions in this field of research. Questions which are usually addressed in market research are as the following: What is the market size? How is the market changing? What is the future of the market? How is the supply chain to be planned? How is manufacturing to be planned? What kind of promotion is to be offered? When is the promotion to be offered? To whom is the promotion to be offered? How is the strategy of the organization to be defined? What is the preference of the customers? What are the real needs of the customers? What is the competition? Where is the opportunity? What is the target market of the product? What are the market segmentations? What is the success factor in each market segmentation?

Models used in market research can be divided into the two categories of primary research and secondary research based on the data utilized. In primary research models, the researcher will design and conduct surveys, questionnaires, and interviews to capture required data for the analysis. In secondary research models, the researcher will use data captured from other sources, such as online data or published data in research papers.

## 2.5. Critiques of Current Models

All current models that can be used for predicting acceptance of a new technology are unreliable and have unstable prediction power (Bento, Li, and Roth, 2012). This is partly

due to the nature of humans. Human behavior is difficult to predict since there are so many unknown factors involved. Further the decision-making process of each individual varies and is not completely known (Bento, Kenneth Gillingham, and Karen Palmer, 2014). The difficulty in developing an accurate prediction model to consider diversity of customers and heterogeneity of their preferences has been highlighted by Howarth and Sanstad (1995) and Bento, Li, and Roth (2012) without a proper solution. In addition, current econometric models are based on assumptions that are not valid (Kahneman and Tversky, 1979). For instance, many customers do not have enough knowledge about the characteristics of the innovative technologies. Also, the majority of them do not know how to calculate the present value of future gains or the utility function. Even if they know, individuals do not perform these calculations as part of their decision-making process (Kahneman and Tversky, 1979) and simplify decisions by considering only a subset of the available information (Simon, 1955). The deficiency of assumptions in econometric models is highlighted by Savage (1954) as well. He showed that current econometric models, including Discrete Choice, have little or no predictive power outside of their bounded domains since predicted rational decisions based on Utility Theory occur only under some conditions (Warren and Simpson, 1976). Later, Turrentine and Kurani (2007) developed a semi-structured interview which was taken by 57 households in a 12-month period. The study covered nine different lifestyles for acceptance of innovative efficient technologies. These researchers questioned econometric modeling and showed that individuals make decisions in a very simple way and do not engage in calculated decision making. Customers look for heuristic shortcuts for decision making. Even the Energy Efficiency Gap, which was introduced to help understand the gap between real acceptance of new energy efficient products and predicted acceptance by econometric models, fails (Howarth and Sanstad, 1995). Many researchers tried to evaluate the existence of the Energy Efficiency Gap by calculating the implied discount rate, including Wolverton (2011) and Gallego et al. (2013), but still a conflict exists. Wolverton (2011) and Train (1985) showed a high implied discount rate, and Metcalf (1999) and Gallego et al. (2013) showed a lower implied discount rate. Gilbert E. Metcalf and Kevin A. Hassett (1993) and Sutherland (1991) who believe a high implied discount exists, concluded so as the result of customers' rational thinking.

*Table 2.1: Critiques of Current Models*

| Year | Author | Summary | Criticize |
|---|---|---|---|
| 1954 | Savage | Econometric models have little or no predictive power outside of their bounded domains | Econometric Models |
| 1974 | Simpson | Rational decisions based on utility theory occur only under some conditions | Econometric Models |
| 1979 | Kahneman and Tversky | Indication that customers decisions are violating the rational choice | Econometric Models, Rational Choice Theory |
| 2007 | Turrentine and Kurani | Individuals make decision in a very simple way and do not make calculations | Econometric Models |
| 2002 2011 | Kahneman | Customers do not use a rational calculating process. Instead, they use simple heuristic methods or make decisions under the influence of emotion and image | Econometric Models |
| 1993 | Gilbert E. Metcalf and Kevin A. Hassett | Validated econometric models and concluded that high expected rate of return is the result of their rational thinking | Energy Efficiency Gap |
| 1995 | Howarth, R B and Sanstad, A H | Reviewed econometric models and suggested to observe the customers' actual decision instead since models fail to find the discount rate in energy efficiency technologies | Energy Efficiency Gap and Econometric Models |
| 2003 | Legris, Ingham, and Collerette | TAM account for only 40% of a technological system's use | TAM |
| 2012 | Bento, Li, and Roth | Models are biased since they do not consider heterogeneity | All current models |

They highlighted that customers are uncertain if their investment in a more expensive energy efficiency technology will pay off and as the result they require a rate of return higher than the market discount rate. This conclusion is questioned by Kempton and Montgomery (1982). He used a simple survey to study the choices of customers facing future savings in energy by more efficient technologies. These authors concluded that customers calculate the future energy saving by using current energy prices at the time of purchase, rather than the future price. Thus, customers ignore future increases in fuel prices at the time of purchasing a new product. In addition, Kempton et al. (1992) showed that customers are more sensitive to the price of a product than they are to saving money on energy in the future. While Energy Efficiency Gap assumes a relation between the future price of energy and the price a customer is willing to pay for an energy efficiency technology, Friedman (2002) questioned this relation and showed that the higher price of energy will motivate customers to consume less rather than motivating them to shift to efficiency technology. Later, TAM considered the effect of customers' image and perceived view regarding the technology as the predictors of the acceptance. Legris, Ingham, and Collerette (2003) studied this model and concluded that TAM accounts for only 40% of a technological system's use. As Kahneman (2011) mentioned, not one of the current models considers customers' simple heuristic decision processes, while ETAM does. Table 2.1 shows the summary of the most important critiques and models or theories which they criticize.

## 2.6. Empirically Studied Attributes

In this section, empirical literatures are summarized into a list of attributes which have already been studied for their effects on acceptance of new energy efficiency technologies. Some of these factors have already been highlighted in the previous sections of this chapter, intermingled among the review of econometric and non-econometric models and theories used in predicting acceptance of new technologies.

Hassett, Metcalf (1995), and Jaffe et al. (1995) showed that increases in energy prices affect the adoption of energy efficiency technologies in a positive way. Later, Yizao Liu

(2014) studied the interaction between where customers live and energy price with the market. He concluded that the effect of energy costs on consumers' preference for an energy efficiency technology is positive and that living in an urban or suburban area increases the possibility of buying an innovative efficient technology. This study developed a customer utility function considering the effect of income on acceptance. See Equation 2.6.

$$U_{nj} = \alpha_n p_j + \beta_n x_{nj} + \gamma_n Tch_j + \sum_{g=1,2,3,4,5} \lambda_g d_{gn} Tch_j + \epsilon_{nj} \quad (2.6)$$

$g \in \{1,2,3,4,5\}$ Income group of customers

$d_{gn}$ Dummy variable identifying customer n as belonging to income group g

$p_j$ Price of the $j^{th}$ good

$Tch_j$ Dummy variable indicating an innovative energy efficient good

$U_{nj}$ Utility Function from the $j^{th}$ choice for the $n^{th}$ customer

$x_{nj}$ Vector of observed characteristics of $j^{th}$ choice by $n^{th}$ consumer

$\epsilon_{nj}$ Unobserved random error

If customer n faces a choice among j goods, he chooses the one that maximizes the utility function. Coefficients $\alpha_n, \beta_n$ and $\gamma_n$ are assumed to vary among customers. The preference of innovative efficiency technology in the lowest income group is $\gamma_n$ and for higher income groups is $\gamma_n + \lambda_g$. $Tch_j$ is a dummy variable which is equal to 1 if the choice is an innovative energy efficient product, and is equal to 0 otherwise. The preference for innovative efficient technology is specified to vary across different groups, g, with the lowest in Group 1. Individuals in Group 1 have an income lower than $25,000 per year. Individuals in Group 2 has an income equal to or greater than $25,000 and equal to or less than $49,999 per year. Individuals in Group 3 have an income equal to or greater than $50,000 and equal to or less than $75,999 per year. Individuals in Group 4 have an income that is equal to or greater than $75,000 and equal to or less than $99,000 per year. Individuals in Group 5 have an income greater than $100,000 per year.

Alan Jenn et al. (2013) studied the effect of the Energy Policy Act of 2005 on expanding the market share of innovative efficiency technologies and showed that it was a positive effect. In addition, Hyundo Choia and Inha Ohb (2010) used surveys and conjoint analysis to study the effect of policy on the sales of innovative efficient technologies and they confirmed the result of previous studies. On the other hand, Stern (1985) showed that incentives to accept energy efficiency products are not as effective as was assumed, which is in contrast with Alan Jenn (2013), Hyundo Choia and Inha Ohb (2010). Gillingham, Newell, and Palmer (2009) considered other attributes besides policy and price. According to them, intensity of use of the product, equipment lifetime, environmental concerns, lack of information, and policy are important factors in accepting the innovative efficient technologies. They also mentioned the importance of Learning by Doing (LBD). According to Arrow (1962), the concept of LBD means that by increasing the amount of production of a new technology, manufacturers learn how to reduce the price and increase the quality. This positively impacts the market.

Sanstad et al. (2006) and Jaffe et al. (2004) showed the effect of information regarding the positive outcomes of choosing an efficient technology, including savings from energy cost and incentives, is an important factor. In addition, Jaffe and Stavins (1994) showed the negative effect of incomplete information in undervaluing future saving of efficient technologies by customers. While importance of information about characteristics and performance of technologies is believed to be significant, Carpenter and Chester (1984) showed that information is not as important as presented in other studies. They ran a survey about tax credits in the early 1980's for reducing energy consumption and found that although 86% of individuals were aware of the incentives, only 35% used the offered incentives. This can be the result of not considering customers' credit scores. Berry (1984) showed that customers who can borrow at a lower interest rate are more willing to invest in energy efficiency products. According to Schultz, Khazian, and Zaleski (2008) in addition to information and communication, social norms have an effect on accepting the new efficiency technologies. Table 2.2 shows a list of attributes considered in previous models as possible important factors in accepting innovative efficient technologies.

*Table 2.2: Attributes Considered in Previous Models*

| Year | Author | Attributes Considered in the Study |
|---|---|---|
| 1962 | Rogers | Invention, time, communication channels, social system (social status, education and income) |
| 1966 | Lancaster | Invention characteristics |
| 1976 | McFadden | Invention characteristics, customer characteristics |
| 1989 | Fred D. Davis | Viewpoint and image |
| 1980 | Blumstein, C., et al. | Social norms, interest rate, policy and regulation, income, information |
| 1979 | Hausman | Usage, policy and regulation, information, misplaced incentives, attitude toward energy efficiency, access to financial resources, energy price |
| 2006 | Sanstad et al. | Information |
| 2009 | Gillingham, Newell, and Palmer | Energy prices, intensity of use, equipment lifetime, environmental, lack of information, policy |
| 2013 | Kenneth Gillingham and Karen Palmer | Credit constraints, regulatory failures, preferences, habits |
| 2013 | Jenn et al. | Energy policy by government |
| 2014 | Yizao Liu | Income, choice alternatives (product characteristics), price of the technology |

# Chapter 3: Methodology

Methodology consists of four major parts: selection of input attributes, data collection, data analysis, and model evaluation.

Differences in individual characteristics, individual environments, and technology process states, which may result in different individual decision processes, will be inputs of the prediction part of ETAM. To help selecting the attributes, a novel breakdown of attributes will be introduced in the first section of this chapter. In the second section of this chapter, data collection in ETAM is presented. In the third section, prediction in ETAM is illustrated and the fourth section will discuss the validation of the model. Figure 3.1 shows conceptual framework of ETAM.

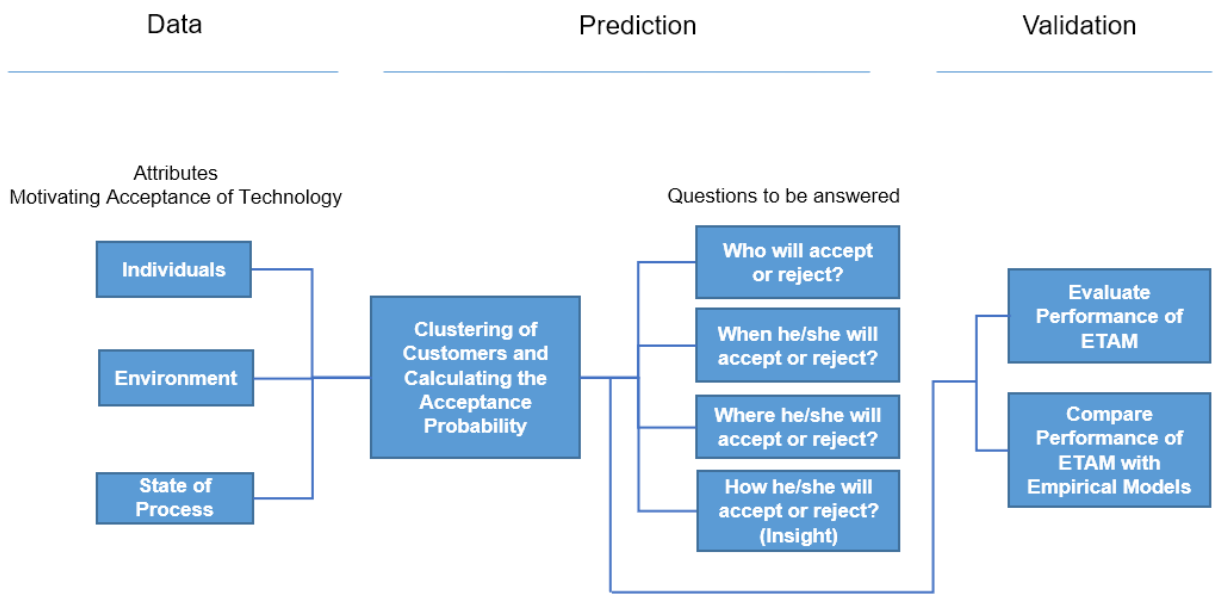



*Figure 3.1: ETAM Conceptual Framework*

The analysis part of the ETAM overcomes the assumptions in previous models considering the critiques discussed in Section 2.5. In contrast to previous models, ETAM does not use the Utility and Rational Choice theories. It does not assume the customers

to be informed Economy Rational individuals who know characteristics of the product and know how to make calculations to maximize their gain. ETAM assumes that customers use simple heuristic decision processes by answering questions in their own minds, which may result in rational or non-rational choices. It clusters individuals and considers their heterogeneity. ETAM is a parametric and probabilistic model which not only predicts, but also describes the acceptance of new technologies by individuals. ETAM will answer questions of when, where, and how technology acceptance will be achieved and by whom.

## 3.1. Selection of Attributes

The limited number of studied factors in previous research has been inadequate to understand the complicated behavior of customers and their assumptions. The majority of attributes which have been introduced as significant ones in acceptance of innovative technology are correlated (Bento, Li, and Roth, 2012). Examples include income and environmental attributes. Models evaluated in the literature study did not compare all attributes to select the best ones for customer prediction. There is a demand to identify attributes that have greater significance for predicting customers.

This study proposes a novel comprehensive breakdown of probable significant attributes to be used for selection of input attributes of the proposed model based on the literature study.

Figure 3.2 shows the breakdown of the groups of attributes and their connection with previous studies. Some groups of attributes have been considered as the input of the known econometric and non-econometric models. These are identified by a “Yes.” marked in Figure 3.2. The researcher who studied the specific set of attributes is mentioned in the last column.

The breakdown in Figure 3.2 consists of three levels. The first level contains the primary categories of attributes. To articulate each of these primary categories, they have been broken down into one or two more detailed subcategories which are referred to as second and third level categories of attributes. The final subcategories from the primary categories contain attributes which do not narrow down further into another level. The

final subcategories of attributes may be second level attributes or third level attributes. Each of the 18 final subcategories of attributes in ETAM is described by a number of attributes. As shown in the figure, actual usage information of customers has not been used as the input of empirical models for predicting the acceptance of innovative technologies. Having all categories of previously studied attributes in addition to this new category of attributes helps to compare their significance and understand which of them are the best for prediction. The following will illustrate the three levels of the breakdown and an example of attributes describing the final subcategories.

### 3.1.1. First Level of the Breakdown

According to the first level of the comprehensive breakdown of attributes in ETAM, Figure 3.2, attributes which may motivate acceptance or rejection of a new technology describe Individuals, the Environment where individuals reside, and the State of Process of the technology, as indicated by Rogers (1962). Attributes may exist which describe more than one of the above three primary categories. Attributes which describe more than one category may be considered only once by using the union in set theory in the process of data collection to reduce the amount of data collection. See Equation 3.1.

$$\mathrm{M} = (\mathrm{I} \cup \mathrm{E} \cup \mathrm{S}) - (\mathrm{I} \cap \mathrm{E}) - (\mathrm{E} \cap \mathrm{S}) - (\mathrm{I} \cap \mathrm{S}) + (\mathrm{I} \cap \mathrm{E} \cap \mathrm{S}) \quad (3.1)$$

$\mathrm{E}$ Attributes that describe Environment

$\mathrm{I}$ Attributes that describe Individuals

$\mathrm{S}$ Attributes that describe the State of Process

$\mathrm{M}$ Attributes which may motivate acceptance or rejection of a new technology

$(\mathrm{I} \cup \mathrm{E} \cup \mathrm{S})$ is the union of all attributes. $(\mathrm{I} \cap \mathrm{E})$, $(\mathrm{E} \cap \mathrm{S})$, and $(\mathrm{I} \cap \mathrm{S})$ are the attributes which describe two primary categories while $(\mathrm{I} \cap \mathrm{E} \cap \mathrm{S})$ are attributes which describe all three primary categories, such as country.

| Comprehensive Break Down of Attributes | | Diffusion by Rogers 1962 | Classical Theory of Customer Demand 1900 | New Theory of Customer Demand by Lancaster 1966 | Random Utility Theory, Discret Choice Model by McFadden 1976 | Technology Acceptance Model by Davis 1989 | Other Studies | ETAM by Ali 2018 |
|---|---|---|---|---|---|---|---|---|
| 1 | Individual | | | | | | | |
| 1.1 | Social | | | | | | | |
| 1.1.1 | Demographic | Yes | | | Yes | | Blumstain et al. (1980) | Yes |
| 1.1.2 | Socio-Economic | Yes | | | Yes | | Hausman (1979), Kenneth Gillingham and Karen Palmer (2013), Yizao Liu (2014) | Yes |
| 1.1.3 | Occupation | Yes | | | | | Kenneth Gillingham and Karen Palmer (2013) | Yes |
| 1.1.4 | Education | Yes | | | | | | Yes |
| 1.1.5 | Habits | | | | | | Blumstain et al. (1980), Kenneth Gillingham and Karen Palmer (2013) | Yes |
| 1.1.6 | Beliefs and Values | | | | | Yes | | Yes |
| 1.2 | Knowledge | | | | | | | |
| 1.2.1 | Training | | | | | | Hausman (1979), Sanstad et al. (2006), Gillingham, Newell and Palmer (2006) | Yes |
| 1.2.2 | General Knowledge through Media | Yes | | | | | Hausman (1979), Gillingham, Newell and Palmer (2006) | Yes |
| 1.3 | Intended Use of Innovation | | | | | | | |
| 1.3.1 | Type of Usage | | | | | | | Yes |
| 1.3.2 | Amount of Usage | | Yes | | | | Hausman (1979), Gillingham, Newell and Palmer (2006) | Yes |
| 1.3.3 | Energy Unit Cost | | Yes | | | | Hausman (1979), Gillingham, Newell and Palmer (2006) | Yes |
| 2 | Environment | | | | | | | |
| 2.1 | Geographic | | | | | | | |
| 2.1.1 | Population Work and Wealth Status | Yes | | | | | Gillingham, Newell and Palmer (2006), Jenn et al. (2013) | Yes |
| 2.1.2 | Urban/Rural | Yes | | | | | Gillingham, Newell and Palmer (2006), Jenn et al. (2013) | Yes |
| 2.2 | Policies, Standards, and Laws | | | | | | | |
| 2.2.1 | Federal | | | | | | Hausman (1979), Gillingham, Newell and Palmer (2006), Kenneth Gillingham and Karen Palmer (2013) | Yes |
| 2.2.2 | State | | | | | | Hausman (1979), Gillingham, Newell and Palmer (2006), Kenneth Gillingham and Karen Palmer (2013) | Yes |
| 2.2.3 | Business | Yes | | | | | | Yes |
| 3 | State of Process | | | | | | | |
| 3.1 | Diversity of products | Yes | | Yes | Yes | | Yizao Liu (2014) | Yes |
| 3.2 | Market Share | Yes | Yes | | | | Yizao Liu (2014) | Yes |

*Figure 3.2: Breakdown of Attributes*

### 3.1.2. Second Level of the Breakdown

The second level of the breakdown articulates the first level in more detail.

The primary category of attributes describing the characteristics of Individuals consists of three second level categories of Social attributes, Knowledge attributes, and Intended Use of Innovation attributes.

The primary category of attributes describing the characteristics of Environment consists of two second level categories of Geographic attributes and Policy, Standards, and Laws attributes.

The primary category of attributes describing the characteristics of State of Process consists of two second level categories of Diversity of Product attributes and Market Share attributes. This primary category of attributes does not narrow down further to the third level and is consequently discussed in more detail in this section.

Diversity of Products is described by attributes illustrating the variety of options a customer faces when making a choice. According to the New Theory of Consumer Demand by Lancaster (1966), customers buy products for their characteristics. A customer not only considers the innovative technology used in a product but also other characteristics of the product. For example, the number of available products with different characteristics using a specific technology may be an important attribute in the acceptance of that technology.

The importance of considering the market share of the innovative technology as the input of the model can be discussed from three perspectives, outlined as follows.

a) According to the Theory of Diffusion by Rogers (1962), more and more customers will gradually accept a new technology, and in each stage of this process the characteristics of those who are willing to accept the technology are different. This shows the importance of having the market share as an input of the model.

b) Institutional learning is the amount of learning by organizations that affects the final cost of producing a unit of a technology. Industries gradually learn more about how to reduce the production cost while they produce the product. That is, they learn by doing. The effect of the final price of a product on its acceptance has always been unneglectable. The Classical Theory of Consumer Demand is based on the fact that when the price of a product decreases, the market share will increase. Kempton et al. (1992) showed that

customers are more sensitive to the sale price of the product than they are to savings on the price of energy in the future.

c) According to Economies of Scale in microeconomics, the increase in the amount of production will decrease the final cost of products. Increase in the market share and production amount will decrease the production cost, which again will motivate increase in sales and market share and further decrease the price. This is the third reason for the importance of considering the market share as an input of the model.

### 3.1.3. Third Level of the Breakdown

The third level of the breakdown reveals 16 detailed categories of attributes to further narrow down five of the seven categories introduced in the second level (Figure 3.2). Attributes that are grouped in these 16 third level categories and in the two final categories at the second level are to be used for ETAM data collection.

The second level category of Social attributes consists of third level Demographic attributes, Socio-Economic attributes, Occupation attributes, Education attributes, Habits attributes, and Beliefs and Values attributes.

The third level category of Demographic attributes includes information about race, sex, medical condition, number of household members, and any other attribute regarding the demographic of individuals. The importance of these attributes in the acceptance of innovative technologies have been studied by Blumstein, C., et al. (1980). Also, these factors are considered when applying McFadden's (1976) Discrete Choice Model and the Theory of Diffusion by Rogers (1962). Demographic and cultural factors affect preferences of individuals and their resistance against new ideas and technologies.

The third level category of Socio-Economic attributes includes information which represents the economic situation of the customer and the household he lives in, factors such as income, house ownership, and count of vehicles owned. There are many other factors that can fit in this category and show the amount of wealth or income an individual has. Many researchers studied the effect of financial status and credit rating of individuals on the acceptance of new efficiency technologies and their willingness to spend a higher premium for future saving in energy. These researchers include Hausman (1979),

Kenneth Gillingham, Karen Palmer (2014), and Yizao Liu (2014). Also, the Theory of Diffusion by Rogers (1962) indicates income and wealth as important factors for the amount of risk individuals are willing to take which will place them in one of the categories of innovators, early adopters, early majority, late majority, and laggard adopters. Innovators are the first group to accept a technology, and laggards are the last, as discussed in Chapter 2.

The third level category of Occupation attributes includes information regarding individual occupation details, such as employed or not, employed part-time or full-time, employed to work from home or not, and self-employed or not. There are many other characteristics of the individual occupation that can fit here. The Theory of Diffusion by Rogers (1962) considers occupation to be an important factor in acceptance of efficiency technologies because it is linked to social status, which helps in understanding which group of adapters an individual belongs to. Kenneth Gillingham, and Karen Palmer (2014) reiterated the significance of occupation in predicting the acceptance of efficiency technologies.

The third level category of Education attributes includes information such as highest attained level of education, major, school or university, state, and other related attributes describing the education status of the individual. Education is considered an important factor in defining social status, which is important in understanding acceptance of new technologies (Rogers, 1962). However, education has not been studied as widely as many other attributes including demographic, social, and occupation attributes for its effect on acceptance of new efficiency technologies.

The third level category of Habits attributes describes the habitual behavior of the individual, and it includes information about technology usage, length or time period for which a product is kept, and where and when a new technology purchase is made. The importance of individual habits on the acceptance of innovative energy efficiency technology has been considered by Blumstein, C, et al. (1980), Kenneth Gillingham, and Karen Palmer (2014).

The third level category of Beliefs and Values attributes consists of the answers given by the individual reflecting his or her beliefs, concerns, and perceived view regarding the technology, its usefulness, and what it is going to be used for. For example, safety concerns regarding use of the technology or belief about the amount of energy consumed

by the technology might be included. This type of attribute has not been considered in econometric models because they assume customers to be informed Economy Rational buyers who are not biased and who pick the choice that maximizes their utility function. Davis (1989) questioned this assumption and used perceived views, which are called beliefs in this research, to develop a new prediction model known as TAM. This model showed the importance of these attributes in predicting the acceptance of innovative technologies.

The second level category of Knowledge attributes consists of third level Training attributes and General Knowledge through Media attributes.

The third level category of Training attributes includes information that shows any voluntarily or non-voluntary training which provided information regarding the innovative technology or its alternatives, including weaknesses, strengths, and usage or maintenance information. Econometric models assume consumers to be informed individuals. This means that they have basic knowledge of math and the advantages of technologies in order to make calculations and determine paybacks. Also, individuals who will need to change their behavior and start using a new technology may resist against the change since they have to learn how the new technology works. The importance of training can be seen in many other previous studies including Hausman (1979), Sanstad et al. (2006), Gillingham, Newell, and Palmer (2006).

The third level category of General Knowledge through Media attributes represents the amount of general information individuals receive through media and the type of media used; for example, the amount of internet usage or amount of time spent watching TV or reading newspapers might be included. General knowledge is like training, but the depth is different. Training is customized for individuals. Knowledge gained through media is less in-depth and is not tailored for a limited audience, as is a training session. General knowledge is an important attribute used by Rogers (1962) to describe why the Theory of Diffusion exists.

The second level category of Intended Use of Innovation attributes consists of the third level Type of Usage attributes, Amount of Usage attributes and Energy Unit Cost attributes.

The third level category of Type of Usage attributes includes information that represents where, when, and for what purpose the technology or its alternative is used by individuals. These attributes are derived by aggregating observed usage of the innovative technology or its alternatives by the individuals. For example, the amount of usage in each day of the week or in different zip codes or for the purpose of usage might be counted by these attributes. While the amount of usage has been considered by many researchers, the actual usage of the technology has not been used as the input of a model to predict acceptance of new efficiency technology. This study considers usage attributes as an input of the model.

The third level category of Amount of Usage attributes includes information that shows how much the technology or its alternative is used by the individuals in a specified length of time. For example, the total number of hours the technology is used by the individuals in a year is considered. The amount of usage and the energy unit cost are the first two factors considered to be important in acceptance of innovative technologies. These factors have been considered in the Classical Theory of Customer Demand for predicting acceptance. According to the model, innovative efficiency technology would not be an Economy Rational choice if the customer does not use the technology enough that future saving covers the initial premium cost. Also, the importance of these two types of factors has been highlighted by Hausman (1979) and Gillingham, Newell, and Palmer (2006).

The third level category of Energy Unit Cost attributes include the price of a unit of energy at the time the individual picked or bought the current technology or product. Unit cost is used in the Classical Theory of Consumer Demand similarly to amount of usage attribute. It is also mentioned in many studies as an important factor in decision making. According to Kempton and Montgomery (1982), customers only consider the energy price at the time of purchase, not the future price, for decision making.

The second level category of Geographic attributes consists of third level Population Work and Wealth Status attributes and Urban/Rural attributes.

The third level category of Population Work and Wealth Status attributes consists of information indicating the demographic and economic situation of the area in which individuals reside, such as income per capita, unemployment rate, and renter percentage in a unit area. According to the Theory of Diffusion by Rogers (1962), many customers

wait until other individuals surrounding them accept the innovative technology before accepting it themselves. As a result, the neighborhood acceptance rate will affect their decision. As discussed earlier, wealth has already been considered as an important factor in acceptance. Also, the importance of environmental factors has been considered by Gillingham, Newell, Palmer (2006), and Jenn et al. (2013).

The third level category of Urban/Rural attributes includes any information that describes the area where the observed individual resides, such as population density, weather, urban or rural location, or the type of transportation used. These factors are important in the Theory of Diffusion by Rogers (1962). Also, the structure of the area where individuals reside affects when and how word of mouth will spread.

The second level category of Policies, Standards, and Laws attributes consists of third level Federal attributes, State attributes, and Business attributes.

The third level category of Federal attributes includes information about any monetary and non-monetary incentives offered by the federal government to motivate acceptance of the new, more efficient technology such as tax returns, non-monetary incentives, standards, and limits on the amount of energy consumption by the products. Many researchers have studied the effect of policy and incentives on acceptance of efficiency technologies including Hausman (1979), Gillingham, Newell, Palmer (2006), Kenneth Gillingham, and Karen Palmer (2013).

The third level category of State attributes includes any monetary and non-monetary incentives offered by the state government to motivate acceptance of the new, more efficient technology such as tax returns, non-monetary incentives, standards, and limits on the amount of energy consumption by the products.

The third level category of Business attributes includes any information about monetary and non-monetary incentives and standards offered or set by an organization or business with which an individual wants to collaborate. These incentives are designed to motivate acceptance of the new, more efficient technology. This also includes standards and business norms set by industries.

Having more descriptive attributes within each final subcategory of attributes helps to increase the accuracy of prediction. Equation 3.2 calculates the total number of attributes from the union of the lowest level of categories in the breakdown using the Inclusion-

Exclusion principle in set theory, which is also known as the Sieve Principal. In this equation, the first term calculates the total number of attributes. The following term completely removes the ones which are counted more than once from the total and adds them back only once.

$$|\bigcup_{j=1}^{18} C_j| = \sum_{j=1}^{18} |C_j| - \sum_{j,k:1\le j\le k\le 18} |C_j \cap C_k| + \sum_{j,k,l:1\le i\le k\le l\le 18} |C_j \cap C_k \cap C_l| - \cdots (-1)^{18-1} |C_1 \cap \ldots \cap C_{18}| \quad (3.2)$$

$C_j$ Vector including attributes which describe the lowest level category, $j$.

## 3.2. Data Collection in ETAM

Data collection is discussed in three sections. The first section illustrates how a comprehensive set of attributes is captured from different resources. The second section discusses how to develop the database of the model. Third section discusses the procedure to validate the database and make sure the database represents the real world.

### 3.2.1. Incorporating a Comprehensive Set of Attributes from Different Sources

The proposed model needs a comprehensive set of attributes to define the final subcategories in the introduced breakdown. A higher number of attributes for describing each category of attributes will result in a higher prediction power from the model. Table 3.1 illustrates sources of data for energy efficiency products and their customers. ETAM collects attributes related to individuals, the energy efficiency market, products, usage of products, policy and government, geography, and environment from these online sources of data. Attributes of interest are the ones which describe the final subcategories of attributes introduced in Section 3.1. Depending on the studied technology, different combinations of available online sources of data can be used. For example, Individual

demographic and economic information can be captured from United States Census Bureau. Market information of energy efficiency products can be captured from the California Center for Sustainable Energy. Product information can be downloaded from manufacturers. Information about energy efficiency incentive programs can be captured from the Internal Revenue Service. Geographic information concerning customers can be attained through United States Department of Agriculture and United States Census Bureau.

*Table 3.1: Sources of Information*

| Source of Information | Description |
|---|---|
| Internal Revenue Service | Includes tax incentive information for efficiency technologies |
| U.S. Energy Information Administration | Includes information regarding energy consumption and cost of energy |
| California Center for Sustainable Energy | Includes information regarding action taken by the state to motivate sustainable energy |
| United States Census Bureau | Includes information regarding the demographics of customers in the US |
| Center for Disease Control and Prevention | Includes US population health information |
| U.S. Department of Agriculture | Includes information regarding environment and agriculture in the US and farmers' choice of implements |
| Oak Ridge National Laboratory | Includes a wide range of information from different areas including energy in the US |
| National Highway Traffic Safety Administration | Includes information related to transportation behavior of individuals and their choice of automotive technology in the US |
| Manufacturers | Includes technology used in customer products |
| Retailers | Includes technology used in customer products |
| Credit Card Companies | Includes income information and energy related costs of individuals |
| Flowingdata | Includes a wide range of energy related information |
| Openstreetmap | Includes maps and geographic information of customers |
| Geocommons | Includes a wide range of energy related information from different countries |

*Table 3.1 Continued*

| Source of Information | Description |
|---|---|
| Google | Includes a wide range of energy and technology related information |
| UNdata | Includes a wide range of energy and technology related information |
| World Health Organization | Includes a wide range of information related to energy, health, disease, and epidemic |
| Organization for Economic Co-Operation and Development | Includes information related to the economy of the US and a few other countries |
| data.gov | Includes a wide range of information from the US including energy and technologies |
| DataSF | Includes a wide range of information from San Francisco including energy and technologies |

### 3.2.2. Developing a Relational Database

Information downloaded from online information sources needs to be related, cleaned, and validated for accuracy. While the downloaded data in ETAM is expected to be huge, the file format of most software including Microsoft Access, Excel, and Word is limited to 2GB or so. Rendering huge files that are even smaller than this is still slow and frustrating. Also, most statistical software have limited tools for data manipulation, aggregation, and relation establishment. A good relational database design prevents redundant and incorrect data being stored and makes it possible to relate and validate big data in a timely manner. Redundant and incorrect data occur when the operator misspells an input of the database or uses different terms to refer to the same things.

ETAM suggests using a relational database to relate and store data. Figure 3.3 shows the flow of information from source of data to the database. In a relational database, data are stored in different tables. Each table consists of rows and columns. Columns may also be referred to as fields. Rows are captured information, and fields are attributes. Each table should have at least one field with unique values for each row of data. This field is called the primary key and can be one of the captured attributes or a new field

named as row number or ID number. The primary key also can be a combination of other fields, which results in unique values for rows such as combination of first name, middle name, last name, and date of birth. To be able to relate two tables, A and B, table B needs to have at least one of the fields in table A which can be the primary key or any other fields from table A. This field in table B is called the foreign key. Without the foreign keys, relating tables would not be possible. Like the primary key, the foreign key can be one field or a combination of many fields. In contrast to the primary key, the foreign key is not required to be unique for all rows. If the foreign key is unique for all rows in both table A and table B, then the relationship is known as one-to-one. If the selected foreign key is unique for all rows in table A but not in table B, then the relationship is known as one-to-many. If the selected foreign key is not unique in either table A or table B, then the relationship is known as many-to-many. In data structure design, the many-to-many relation is considered poor design. It increases redundancy, decreases accuracy, and makes data changes more time consuming. Each table may have more than one set of foreign keys to be related to more than one table.

The relational database of ETAM requires at least 7 tables to store and relate data. Figure 3.4 shows the minimum suggested tables and their relation. For demonstration, two factors describe each final subcategory of attributes introduced in Section 3.1. For example, in the Individual Social and Knowledge table, Factor 1 and 2 describe the final subcategory of Demographic attributes. These two factors can be race and sex. There is no maximum limit for number of attributes describing the final subcategories in ETAM. The primary keys of tables are shown with a key indicator. For example, the primary key of the Individual Social and Knowledge table is the Individual ID. This field can be social security or any unique identifier of observed individuals in the table. The Policies, Standards, Incentives, and Laws table uses a combination of two fields of Area ID and Year as its primary key. Area ID can be the abbreviation of states and Year is the year a law or incentive is in place. The primary key of the Technology Information table is Product ID, which can be the unique barcode on each product. Zip code is an excellent choice for the Geographic ID, which is the primary key of the Geographic table. To prevent many-to-many relationships in the database, all observed usages of technology by individuals

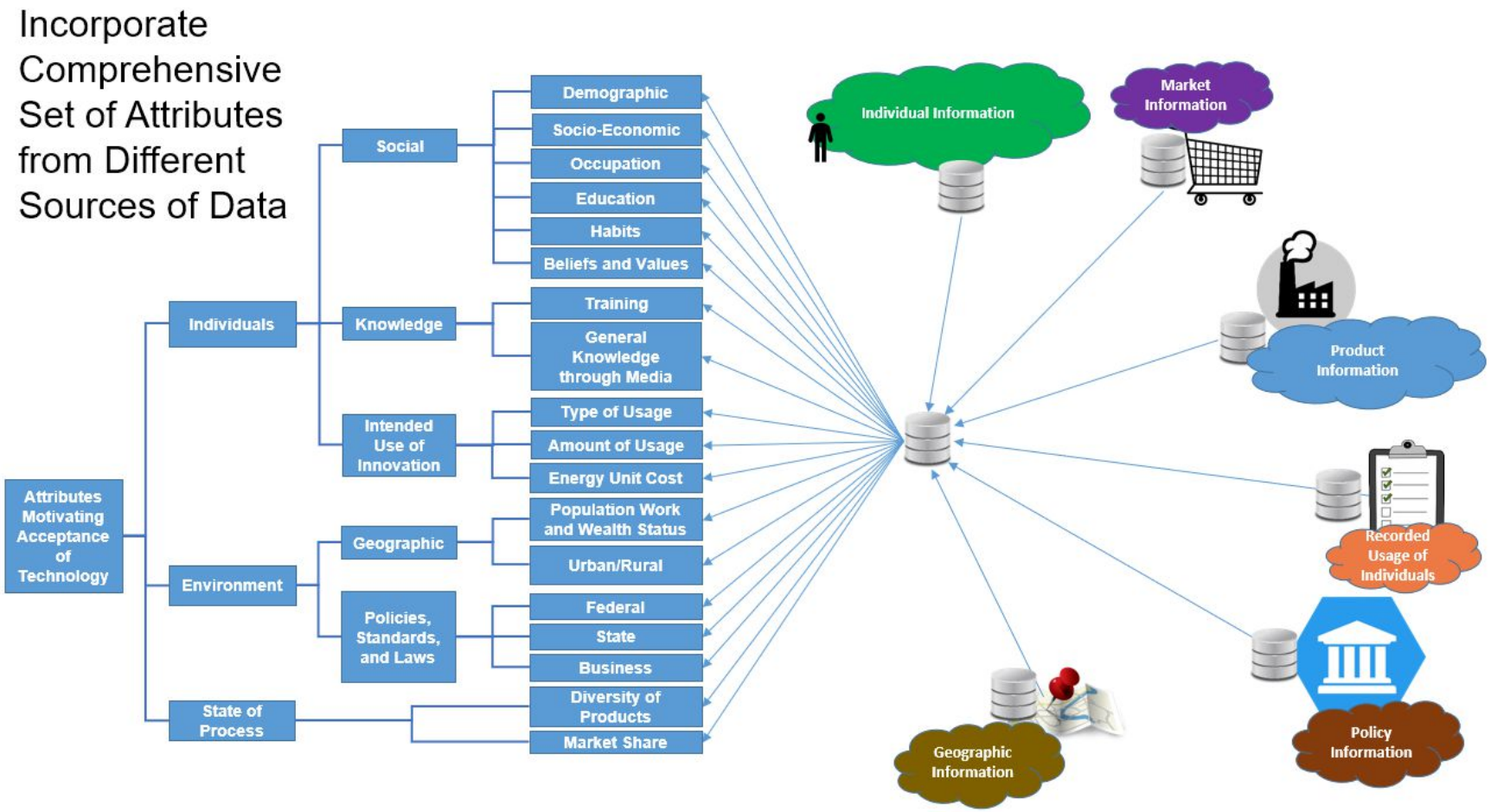



*Figure 3.3: ETAM Data Collection*

are aggregated in the Usage-Aggregated table. This table uses Individual ID as its primary key. The primary key of the Detail of Usage table is a series of sequential numbers shown as Observed Usage ID. The variable of interest to the study is the Type of Technology, and it is stored in the Technology Information table. This field includes the type of technology individuals use or own. Any other field in these tables which is not marked as a primary key, factor, or variable of interest is a foreign key. A primary key may also play as a foreign key. For example, the Individual Social and Knowledge table includes Area ID and Geographic ID (indicating where individuals reside), Product ID (indicating what products individuals own), and Year (indicating when such products were purchased) as foreign keys. Area ID and Year are used as a foreign key to relate this table with the Policy, Standards, and Laws table. The Geographic ID is used to connect this table with the Geographic table. The Individual ID is used to relate this table with the Usage-Aggregated table. Table 3.2 shows the relationships between tables and the foreign keys used to establish them.

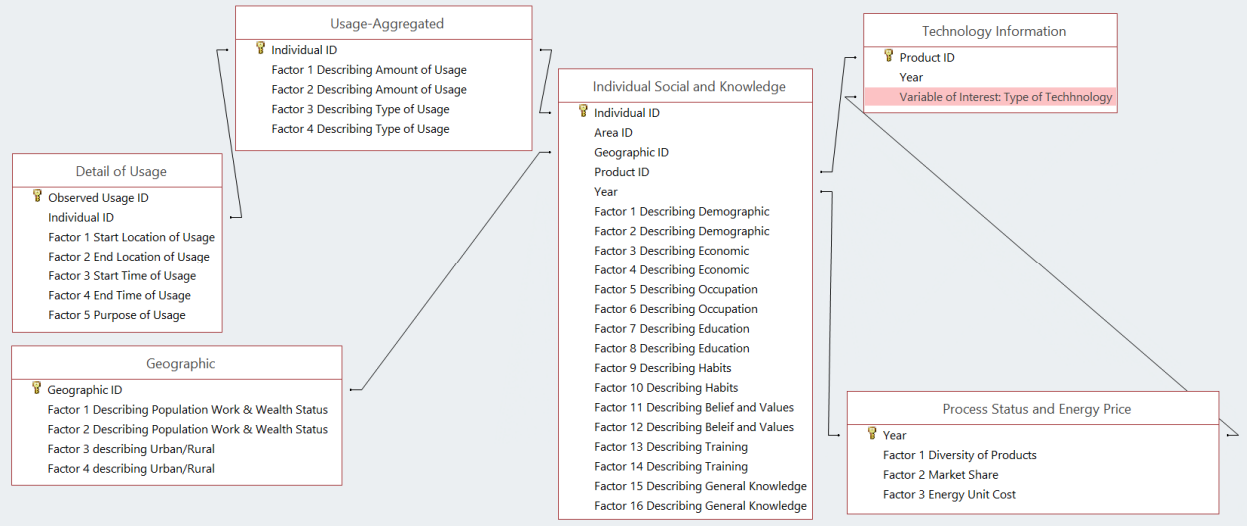



*Figure 3.4: ETAM Relational Database*

*Table 3.2: Relationships of Tables*

| | | |
|---|---|---|
| Individual Social and Knowledge | *Area ID, Year* | Policies, Standards, Incentives, and Laws |
| Individual Social and Knowledge | *Product ID* | Technology Information |
| Individual Social and Knowledge | *Year* | Process Status and Energy Price |
| Individual Social and Knowledge | *Individual ID* | Usage-Aggregated |
| Individual Social and Knowledge | *Geographic ID* | Geographic |
| Technology Information | *Year* | Process Status and Energy Price |
| Detail of Usage | *Individual ID* | Usage-Aggregated |

## 3.2.3. Validating the Database

A database should be validated after it is populated with data. This a critical stage after merging data from various sources of information, especially when sources of information have not been developed for the study. Validation is done in three stages. In the first stage, data is checked for integrity. In the second stage, redundant information is removed, and the third stage will ensure the data in the database represents the real world.

### 3.2.3.1. Ensuring Integrity of Data

Integrity of data ensures data is relevant and is not missing. Integrity of data is achieved through three steps. In the first step, data will be evaluated for relevancy and consistency. In the second step, a method to handle missing data will be introduced. In the third step, the technology used by the observations will be validated.

#### 3.2.3.1.1. Ensure Database is Relevant and Consistent

Any information that is not of interest is removed to ensure data is relevant to and consistent with the interest of the study. For example, we may be interested in the information of individuals who own a house and are between 30 and 50 years old. The database should include only the information of these individuals. Any other information being stored will result in further processes to filter the information. It will also require more hardware resources for data storage and analysis. The study will normally dictate which data are of interest, but there can be other limits by law, geography, or culture. For example, individuals bellow 18 may not be allowed to own a house by law or culture. After defining the scope of data and deleting the ones that are out of scope, the established relationships should be checked for each observation. This means each observation should have values for the attributes.

If some observations are missing values of a few attributes and the number of observations is limited in comparison to the number of attributes, it is possible to manage them properly to prevent losing more information. Managing missing data will be discussed in later sections. However, if all information from a table in the database is missing for an observation, it means that the relation between attributes for that observation could not be established or a huge part of data is missing for that observation. These observations need to be deleted from the database to prevent problems in the analysis part of ETAM. Figure 3.5 shows the algorithm used to ensure relevance and consistency of data in the database.

#### 3.2.3.1.2. Eliminating Missing Data

A part of available data from the National Highway Traffic Safety Administration and the United States Census Bureau consists of questions which have been asked from individuals. Data from other sources of information in Table 3.1 consist of observed or recorded values by operators or machines. Data is recorded in rows of tables in the database. Data is considered to be missing when there is no attribute value for an observation or row of data. This can happen because the person responsible for filling

the information did not enter complete information for a data record or refused to answer a question. Also, recorded answers such as “Do not know” and “Not ascertained” are considered missing data. See Figure 3.6.

Missing data is a well-known problem in the world of data analysis, with no perfect solution. Statistical learning methods (including Regression, Logistic Regression, Time Series, Decision Tree, and Neural Network) have problems with empty fields. If missing data are not handled properly, the result of analysis will not be reliable or may even cause the predictive and descriptive models to fail in finding significant attributes and existing patterns in data. Missing data can be random missing data or non-random missing data. Generally, random missing data can add noise to the analysis, and non- random missing data can result in failure of the model. Missing data can be handled using two different techniques. First, drop the records with missing data. Second, impute missing values and replace them. While dropping the records with missing values looks the easiest and is the most tempting option, in many studies with a limited number of observations, this solution is impractical. This technique neglects the possibility of meaningful trends in missing data. Imputation can be done using various techniques, the following are the well-known ones.

- Mean, median: use the mean or median of the values of other observations of the attribute for the missing one.
- Substitution: substitute the missing value of the attribute with the value obtained from a new observation which previously was not recorded.
- Hot deck: randomly choose the value of the attribute from another individual who has similar values on other attributes to replace the missing one.
- Regression: assume that the attributes with missing values can be predicted by other independent attributes using linear or nonlinear regression.
- Stochastic regression: use the regression method with the addition of a random residual term.

ETAM suggests different solutions for missing data in dependent, categorical independent, and continuous independent variables. The dependent variable, which is also known as the response variable in ETAM, is the field or column of data showing the

ownership of the innovative efficiency technology. This is a binary variable, which means it can have only a value of 1 or 0 for owning or not owning the innovative efficient technology. Other captured variables are independent variables or attributes which will be used to predict the response variable. Figure 3.7 shows the approach used in ETAM for handling the missing data.

It is necessary to ensure that all observations have the information for the response variable. Any record with a missing response variable should be deleted in ETAM unless there is a possibility of capturing it from the observed individual before running the analysis part of the model.

For categorical independent attributes, ETAM considers missing values informative. Missing values are introduced as a new level in each attribute instead of estimating a value for them or dropping them. In other words, a new category is introduced to each categorical attribute, and all missing values are assigned to this category.

For continuous attributes, the observations and records are sorted ascending or descending according to the values of the attribute which has missing values.

Missing values are placed once at the top of the sorted values and once at their bottom. To achieve this, missing values should once receive a value equal to the lowest observed value for the attribute and once receive a value equal to the highest observed value for the attribute. Then, the original column with missing values is dropped and new generated columns of attributes are used as the input of the prediction part of the ETAM. ETAM considers these attributes as two different attributes. Figure 3.8 shows an example of handling missing data by ETAM.

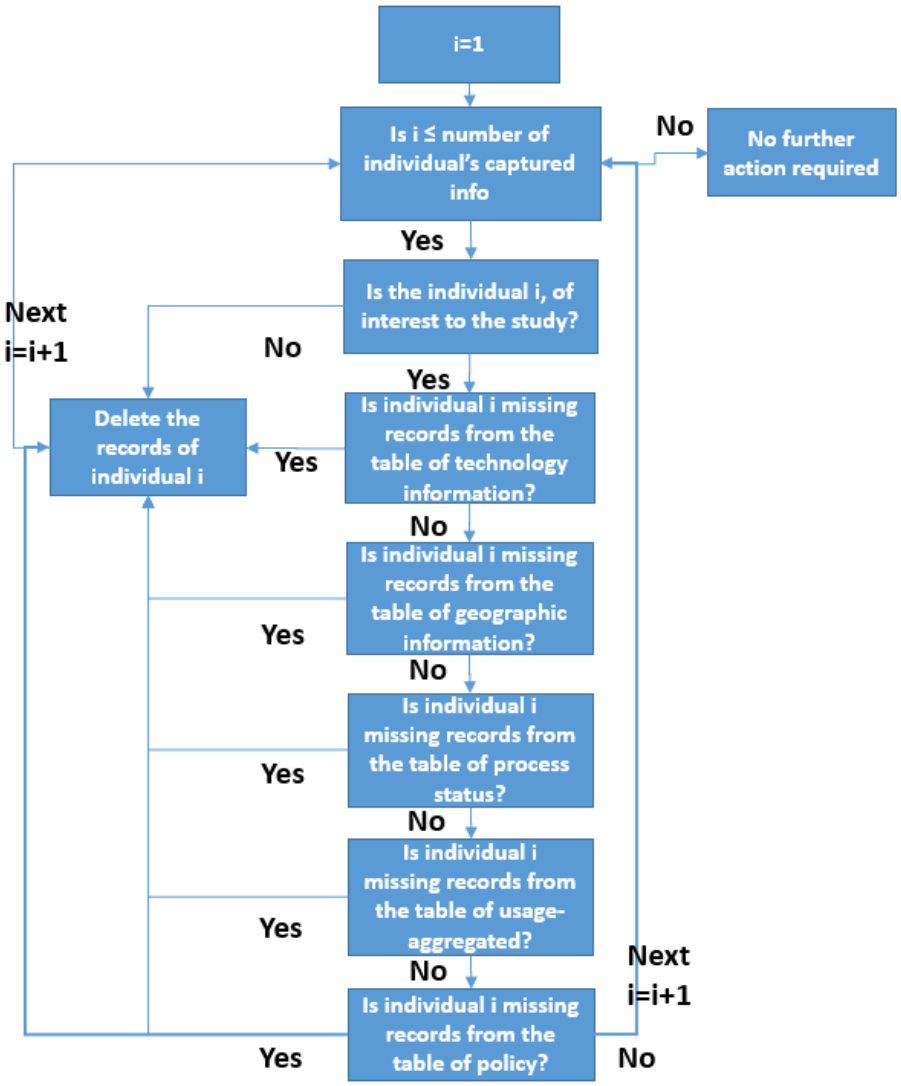



*Figure 3.5: Database Accuracy Algorithm*

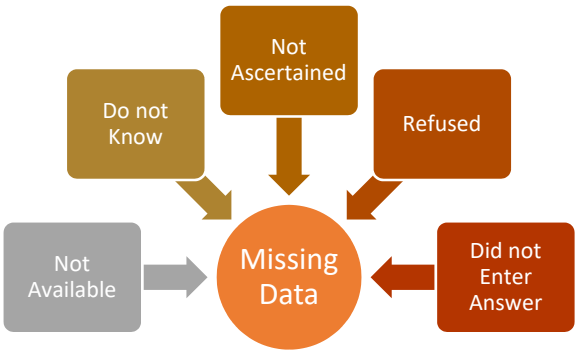



*Figure 3.6: Missing Data*

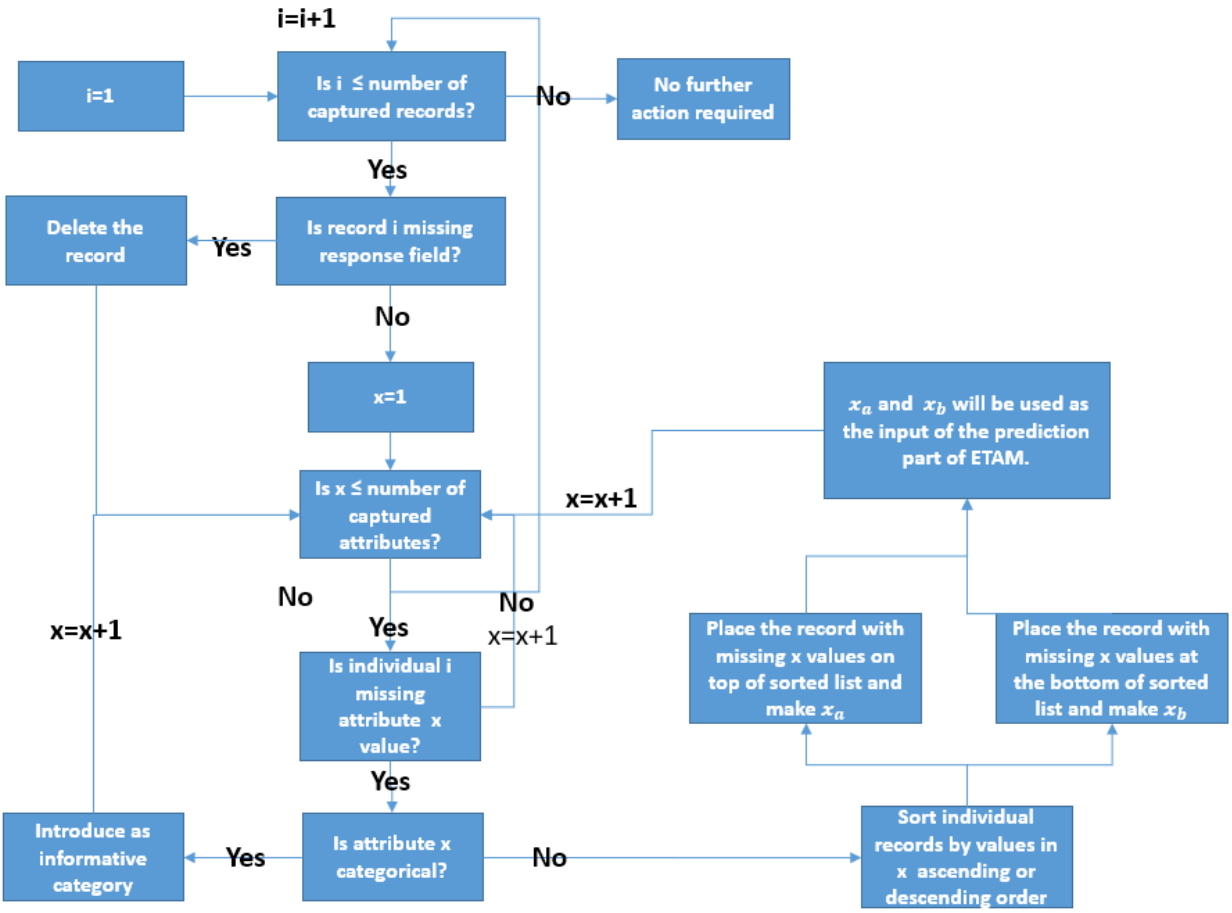



*Figure 3.7: Missing Data Algorithm*

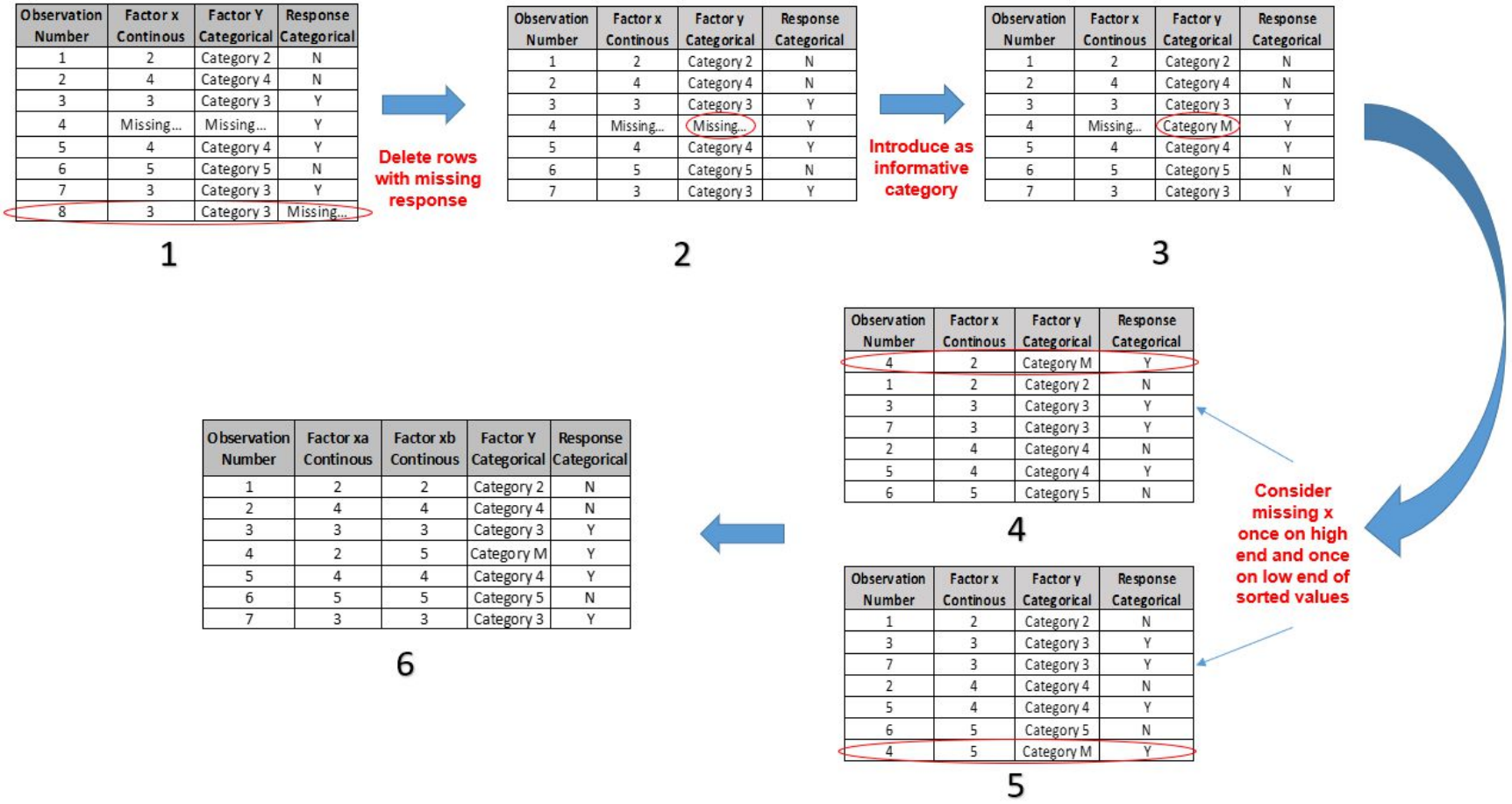



*Figure 3.8: Example of Handling the Missing Data by ETAM*

#### 3.2.3.1.3. Validating Type of Technology Individuals Use

Many individuals are not aware of what technology is used in their purchases. ETAM strongly suggests validating individual responses regarding what technology they use via other sources of information such as manufacturers or retailers. Not doing so may result in failure of the model.

### 3.2.3.2. Removing Redundant Information

Redundant information carries similar information. For example, if the database includes both the unemployment rate and the employment rate of a region, we have redundant information in data. The database should not include redundant attributes. In some cases, redundant attributes may be useful for checking the accuracy of data, but in most cases redundant attributes and fields are considered unnecessary dimensionality in the data. Even if the redundant information is required for validation, such as verifying the type of technology individuals use, only one attribute should be kept after information has already been validated. Unnecessary dimensions of data will increase the processing time, required resources, and (in some statistical techniques) failure of the model to pick the right significant attributes. Figure 3.9 shows the algorithm used to drop redundant attributes.

### 3.2.3.3. Ensuring Database Represents the Real World

Data in the database should represent the real world to prevent biased results in the model. In the process of data collection, if a group of individuals is over sampled or under sampled, the data will not represent the real word anymore. For example, if the ratio of females and males in a studied society is 1:1 but this ratio in captured data is 2:1, then data does not represent that society. Females have been over sampled and males have been under sampled.

Over and under samples are expected when data is pulled from a source of information which has not been developed for the interest of the study. Other factors may also cause non-random samples of the society. For example, when performing a phone interview,

the number of females, males, individuals in a certain age range, and employed individuals responding from home at a specific time of day is different.

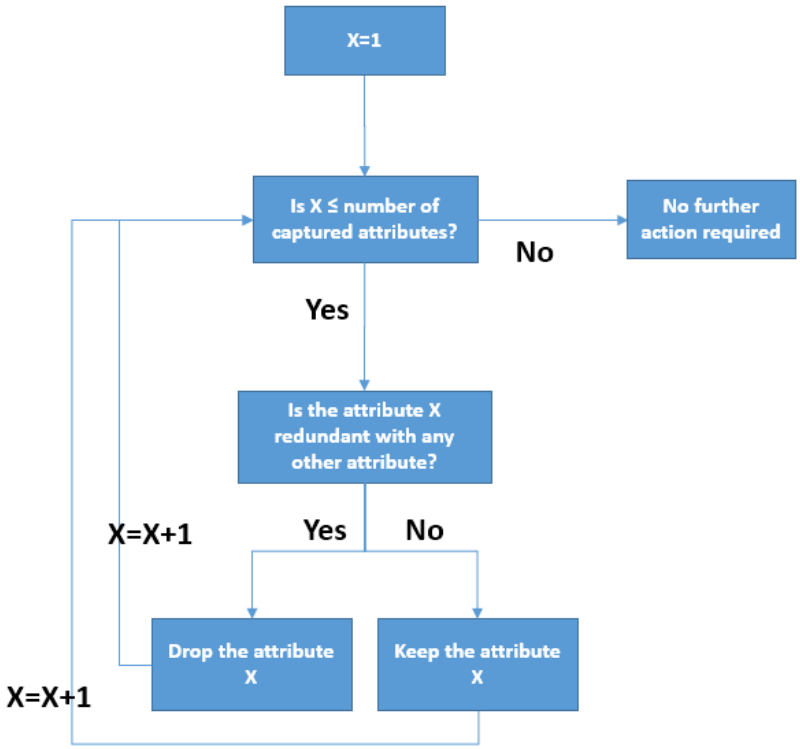



*Figure 3.9: Algorithm to Drop Redundant Attributes*

This changes the probability of talking with a specific group of individuals on the phone, which results in a non-random sample of society. According to Kalton and Graham (1983), using weights to adjust marginal totals of observations which correspond to the target society of study population totals helps to solve this problem. To achieve this, a number of auxiliary variables, such as race and place of residency, are needed. Equation 3.3 shows how the primary weight for an individual is calculated.

$$pw_y^x = \frac{|t_y^x|}{|s_y^x|} \tag{3.3}$$

$pw_y^x$ Primary weight for an individual with a value of y for auxiliary attribute x

$s_y^x$ Subset of sampled data which has a value of y for auxiliary attribute x

$t_y^x$ Subset of target study society which has value of y for auxiliary attribute x

For example, the weight for male observations would be the ratio of men in the target society divided by the ratio of men in captured data. The number of primary weights calculated for each individual is equal to the number of used auxiliary attributes. To calculate the primary weight of each auxiliary attribute for individuals, the distribution of the group in the target society considering the auxiliary attribute, $|t_y^x|$, will be divided by its identical distribution in captured sample data, $|s_y^x|$.

Equation 3.4 illustrates how the primary weights for an individual are made into one weight.

$$wh_i = \prod_X pw_y^x \tag{3.4}$$

$wh_i$ Weight for individual i

The primary weights for individual $i$ are multiplied to make a single weight, $wh_i$ .Each individual has a value of y for each auxiliary attribute x, which results in a primary weight of $pw_y^x$ for individual $i$.

Acceptance of innovative energy efficiency technologies is considered a rare event. Rare events cannot be handled properly by statistical learning tools. Statistical learning tools neglect rare events in favor of other events to reduce the error of the model. Predicting acceptance of innovative efficiency technology is the goal of this study, but it will be neglected by the model if it is not handled properly. This problem can be addressed by oversampling the rare events and undersampling other events or by applying a weight to observations. In ETAM, another multiplier is applied to the previously calculated weight of observations to increase the penalty of neglecting the rare event. See Equation 3.5 and Figure 3.10.

$$FW_i^h = wh_i \times \frac{0.5N}{n_h} \qquad h \in \{0,1\} \tag{3.5}$$

$h \in \{0,1\}$ Dummy variable indicating individual i accepted or rejected efficiency product

$FW_i^h$ Final weight

$n_h$ Number of individuals that accepted or rejected the innovative efficiency product

$N$ Total number of observations

$FW_i^h$ is the final weight after applying the multiplier. The multiplier $\frac{0.5N}{n_h}$ would be different for individuals who accept or reject the innovative technology. $h$ is equal to 0 if individual $i$ rejected the innovative efficiency technology or good, and it is equal to 1 if individual $i$ accepted the innovative efficiency or good. $n_h$ is the number of individuals in observations who accepted, $h = 1$, or rejected , $h = 0$, the innovative product .

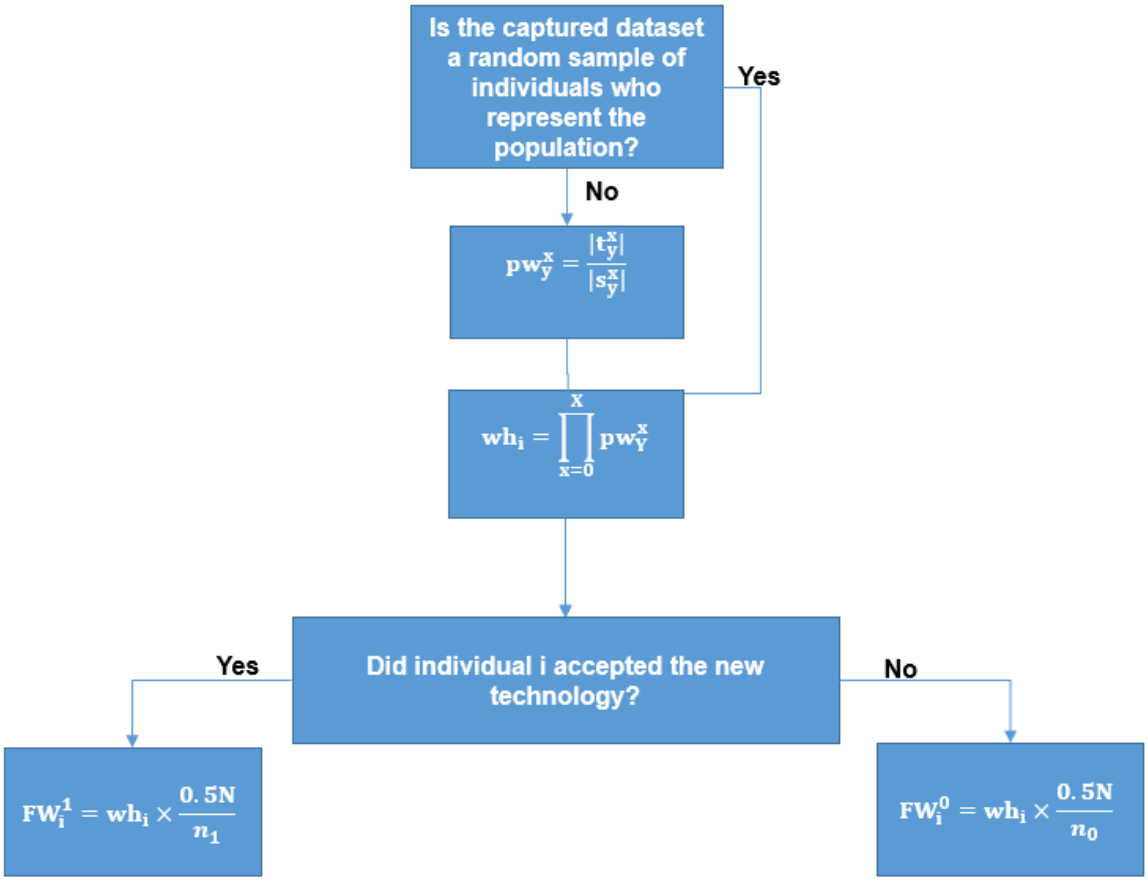



*Figure 3.10: Weight Calculation*

## 3.3. Prediction in ETAM

Development of the prediction part of ETAM is discussed in five sections. The first section illustrates how the collected data is divided into two sets, one for training and one for validating. The second section discusses advantages and details of the suggested statistical learning method. The third section introduces a guideline to assure the accuracy of the prediction model. The fourth section shows how the probability of acceptance by individuals is calculated. The fifth section discusses market opportunities.

### 3.3.1. Dividing Data into Two Sets

With any statistical learning method, it is important to evaluate the result of prediction by a set of data which has not been used for training the model. If the same data which has been used for training the model is used for evaluating the result of prediction, the evaluation cannot be trustable.

ETAM divides captured observations randomly into two sets for training and validating with a ratio of 4:1, as illustrated in Figure 3.11.

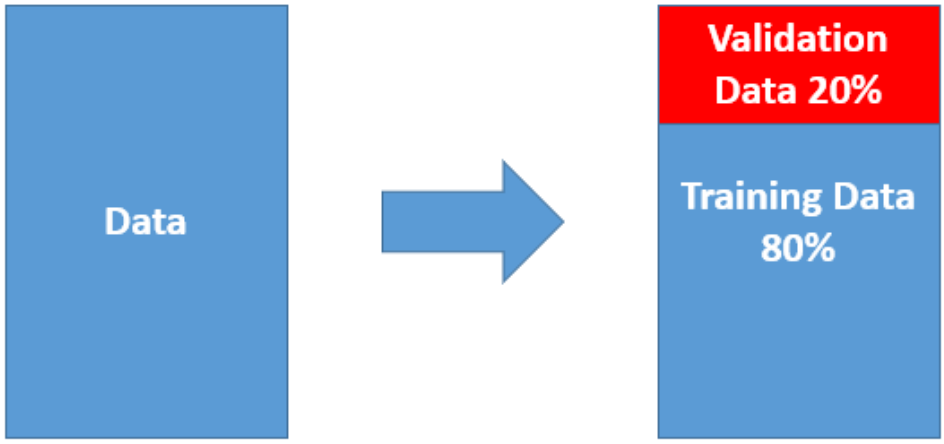



*Figure 3.11: Dividing Data*

### 3.3.2. Simulate Human Decision Processes

The prediction part of the model advances current models by using a supervised clustering method to consider heterogeneity of customers. The model assumes that individuals in different clusters behave differently. Combining the clustering technique with the introduced breakdown of input attributes relaxes current model assumptions regarding informed Economy Rational customers. Individuals may be Economy Rational, do calculations, or just make decisions using heuristic methods and looking at data partially. A decision tree is capable of simulating decision processes of individuals and is the suggested clustering technique for prediction in ETAM. Each node highlights an individually answered question concerning the decision to accept or reject the technology. The model considers the heuristic nature of individual decision-making by being non-parametric in nature and reducing the amount of information important to individuals for decision making in a hierarchal, stepwise order. Clustering is a probabilistic model which can describe and predict. Other considered attributes from the input of the model help to understand which previously studied attributes are really important for prediction and answer the questions of where, when, and how innovative technology is accepted and by whom.

#### 3.3.2.1. Decision Trees for Clustering

A decision tree not only mimics the human process of thinking and decision making, but also has the following advantages over other widely used methods in previous studies for predicting acceptance of new technologies. It considers interaction between input factors by a hierarchy structure. It has no assumptions about linearity and normality of input data. Multicollinearity is not a concern since the decision tree can handle correlated factors and picks the best one for prediction. While data collection in ETAM assures accuracy of data and introduces a technique to consider missing data as informative information, a decision tree is also by nature very robust in tolerating imprecise, conflicting, and missing information. As indicated and in contrast to previous modeling technics, ETAM is capable of modeling complex relations with a lower number of assumptions.

### 3.3.2.2. How the Decision Tree Works in ETAM

The decision tree in ETAM reduces impurity of responses in leaves by splitting observations using independent variables. In other words, observations in child nodes will be purer than their parent node. Figure 3.12 shows a simple example of a decision tree. Here, attributes X and Y are used for splitting. This decision tree results in three purer leaves of individuals, compared to the sampled individuals.

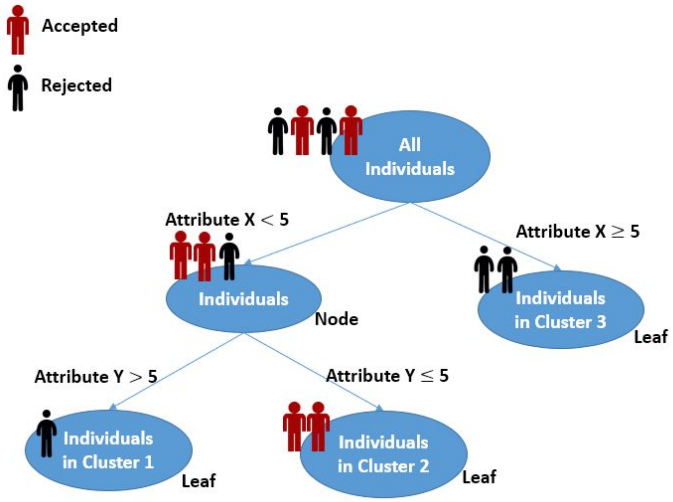



*Figure 3.12: Example of a Decision Tree*

Impurity of data in ETAM is measured by Shannon Entropy. Shannon Entropy measures the average amount of information in each node/leaf. A leaf is the last node which will not be split anymore. The concept of information entropy was introduced for the first time by Claude Shannon (1984). Equation 3.6 illustrates how Shannon Entropy is calculated.

$$\text{Shannon entropy} = -\sum_{i} p_i \log_2 p_i \qquad (3.6)$$

$p_i$ Probability of event i among observations

Shannon Entropy is the sum of the probability of events multiplied by their log.

Events in ETAM are acceptance or rejection of an innovative efficiency product by customers. The data is completely pure when all observations within a node indicate acceptance of the new technology or all indicate rejection of the technology. In such a case the value of the Shannon Entropy will be 0. If half of the observations indicate acceptance of the innovative technology, then the value of Shannon Entropy is 1, which is the maximum possible value. See Figure 3.13.

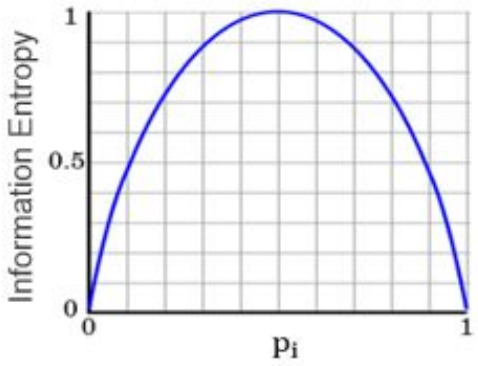



*Figure 3.13: Range of Shannon Entropy*

Many statistical software, including SAS, R, and JMP have the capability of applying a decision tree. If we put all observed values of attributes for individuals from the set of training data in a matrix, then we have the following matrix.

$$\begin{bmatrix} m_1^1 & \cdots & m_1^v \\ \vdots & \ddots & \vdots \\ m_r^1 & \cdots & m_r^v \end{bmatrix} \tag{3.7}$$

$m_r^v$ Observed value of attribute $v$ for individual $r$

Each set of observations for an individual, which is represented as a row of data in the above matrix, will results in acceptance or rejection of the innovative technology. This is written as follows.

$$\begin{bmatrix} m_1^1 & \cdots & m_1^v \\ \vdots & \ddots & \vdots \\ m_r^1 & \cdots & m_r^v \end{bmatrix} \Rightarrow \begin{bmatrix} w_1 \\ \vdots \\ w_r \end{bmatrix} \tag{3.8}$$

$w_r \in \{0,1\}$ Acceptance or rejection of efficiency technology by individual r

$w_r$ is 1 if individual $r$ accepts the innovative technology or 0 if individual $r$ rejects the innovative technology. Figure 3.14 shows the result of splitting observations or individuals by the decision tree using attribute a.

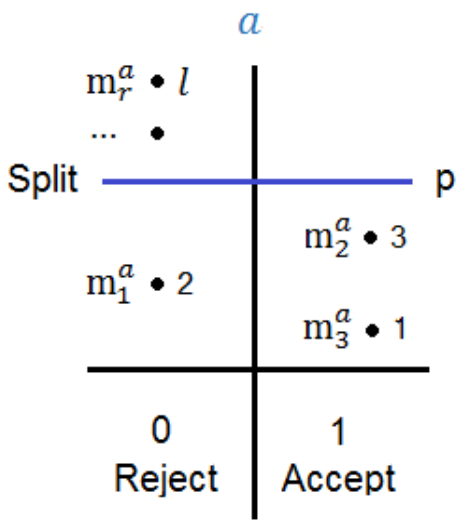



*Figure 3.14: Split in Decision Tree*

This decision tree splits observations or individuals by using attribute $a$ at point p, assuming the response is binary (either acceptance or rejection). A split can also be referred to as a cut. Equation 3.9 shows the entropy after a split at point p using attribute a. The entropy after a split is the weighted average of entropy in child nodes.

$$E(a,p) = \frac{|sl_p^a|}{r}\left(\frac{-\sum_{l=1}^{|sl_p^a|} w_l^a}{|sl_p^a|} \times log_2 \frac{\sum_{l=1}^{|sl_p^a|} w_l^a}{|sl_p^a|} - \frac{|sl_p^a| - \sum_{l=1}^{|sl_p^a|} w_l^a}{|sl_p^a|} \times log_2 \frac{|sl_p^a| - \sum_{l=1}^{|sl_p^a|} w_l^a}{|sl_p^a|}\right.$$
$$+ \frac{r - |sl_p^a|}{r}\left(\frac{-\sum_{|sl_p^a|+1}^{r} w_l^a}{r - |sl_p^a|} \times log_2 \frac{\sum_{|sl_p^a|+1}^{r} w_l^a}{r - |sl_p^a|} - \frac{(r - |sl_p^a|) - \sum_{|sl_p^a|+1}^{r} w_l^a}{r - |sl_p^a|}\right. \quad (3.9)$$
$$\left.\left. \times log_2 \frac{(r - |sl_p^a|) - \sum_{|sl_p^a|+1}^{r} w_l^a}{r - |sl_p^a|}\right)\right)$$

| | |
|---|---|
| a | Selected attribute for splitting |
| $p$ | Value of attribute $a$ used for splitting |
| $sl_p^a$ | Subset of points or individuals below the line $p$ |
| $w_l^a \in \{0,1\}$ | Value of response at point $l$ shown in Figure 3.14 |
| $r$ | Number of observations |

$w_l^a$ is binary and shows the value of response for point $l$ in the surface shown in Figure 3.14. $l$ is an index given to an individual or observation of the matrix shown in Equation 3.7, which is reflected on a surface based on its value of $a$ in Figure 3.14. $sl_p^a$ is a subset of points or individuals in the surface of Figure 3.14 which are located on the lower side of the line $p$. $|sl_p^a|$ is the total number of points in the subset below the line.

The decision tree tries to minimize Equation 3.9 by choosing the best value for p. See Equation 3.10.

To make sure the best attribute, a, is chosen for splitting, the candidate attribute for splitting should achieve the highest amount of gain. Gain is the difference between achieved entropy after a split, which has already been calculated in Equation 3.10, and the entropy of the parent node. See Equation 3.11. Splitting continues in ETAM until the stopping rule is met.

$$\forall a: Min\{\frac{|sl_p^a|}{r}(\frac{-\sum_{l=1}^{|sl_p^a|} w_l^a}{|sl_p^a|} \times log_2 \frac{\sum_{l=1}^{|sl_p^a|} w_l^a}{|sl_p^a|} - \frac{|sl_p^a| - \sum_{l=1}^{|sl_p^a|} w_l^a}{|sl_p^a|} \times log_2 \frac{|sl_p^a| - \sum_{l=1}^{|sl_p^a|} w_l^a}{|sl_p^a|}$$
$$+ \frac{r - |sl_p^a|}{r}(\frac{-\sum_{|sl_p^a|+1}^{r} w_l^a}{r - |sl_p^a|} \times log_2 \frac{\sum_{|sl_p^a|+1}^{r} w_l^a}{r - |sl_p^a|} - \frac{(r - |sl_p^a|) - \sum_{|sl_p^a|+1}^{r} w_l^a}{r - |sl_p^a|} \quad (3.10)$$
$$\times log_2 \frac{(r - |sl_p^a|) - \sum_{|sl_p^a|+1}^{r} w_l^a}{r - |sl_p^a|})\}$$

After a successful split, Equation 3.10 and then 3.11 will be run again at final nodes by considering available observations in each node. The number of observations decreases as the tree grows.

The stopping rule is required to prevent overfitting of the model. The suggested rule is to stop when the number of correct predictions is better than what the next 10 splits would obtain.

$$\mathrm{a}|\mathrm{Max}.[\mathrm{E}(S_j) - \sum_{i \in \{\mathrm{L},R\}} \left(\frac{S_j^i}{S_j}\right) E(S_j^i)] \quad (3.11)$$

$S_j$ Parent Set

$S_j^i$ Subset $S_j$

## 3.3.3. Ensuring Accuracy of the Tree

Having a stopping rule does not guaranty accuracy and reliability of the developed tree. Having more pure leaves is tempting, but a low number of observations in a leaf can be the indication of overfitting and higher errors later when applying the model to new data. Berry and Linoff (1999) suggest 0.25% to 1% of observations as the minimum number of observations in a leaf. Considering 1% as the lowest acceptable number of observations in leaves is more conservative; however, in rare events like energy efficiency technology acceptance, considering 0.25% as the minimum number of observations in leaves may

be a better option. Leaf nodes which include lower than the minimum acceptable number of observations should not be considered as valid clusters when interpreting the results.

### 3.3.4. Defining Probability of Acceptance for Individuals in a Leaf

The probability of acceptance in each leaf is calculated using Bayesian Theory as shown in Equation 3.12. The calculated values show the predicted probability of acceptance by individuals who belong to a cluster or leaf.

$$P_i = P(\text{Acceptance}|\text{Being in Leaf i}) = \frac{P(\text{Being in Leaf i} \cap \text{Acceptance})}{P(\text{Being in Leaf i})} = \frac{|A_i|}{|A_i| + |R_i|} \quad (3.12)$$

$A_i$ Subset of individuals in leaf $i$ who accepted the new technology

$P_i$ Probability of acceptance by individuals who are in leaf or cluster $i$

$R_i$ Subset of individuals in leaf $i$ who rejected the new technology

$|A_i|$ and $|R_i|$ are the total number of individuals in these subsets

### 3.3.5. Evaluating Clusters for Market Opportunity

Manufacturers, retailers, and policy makers are interested in knowing which individuals accept innovative efficient technologies. These are individuals with a higher than average acceptance probability, and they are known as market opportunities in the field of market research. To distinguish these individuals, there is a need to calculate prior probability of acceptance as the indicator of average probability of acceptance. Prior probability of acceptance is the probability of acceptance among individuals in original observed data before any analysis or clustering is applied. This probability is calculated in Equation 3.13.

$$PR = \frac{\text{Total Number of Observed Individuals who Accepted}}{\text{Total Number of Observed Individuals}} \quad (3.13)$$

$PR$ Prior probability of acceptance

Clusters with an acceptance probability higher than the prior probability of acceptance, $PR$, are suggested as the market opportunity by ETAM.

Market opprtunity should be: $P_i > PR$

Figure 3.15 shows the result of a simple decision tree. The decision tree divides the observations into clusters by using cuts parallel to the axes. Each cluster is distinguished by a number of cuts and directions. Axes are the model input attributes selected by Equation 3.11 in a multidimensional page which has possible values between 0 and 10 in this example. The intersections of cuts and axes are defined by Equation 3.10. This simple tree has four leaves, which are indeed clusters of individuals. Each leaf is distinguished by the intersections of these two cuts and two directions. For instance, the cluster in the top right is distinguished by cut a in the increasing direction and cut b in the increasing direction. In this simple tree, four acceptance rates are calculated for clusters. The differences between clusters and their acceptance rates are used to predict customers and answer questions of when, where, and how the innovative energy efficiency technologies are accepted and by whom.

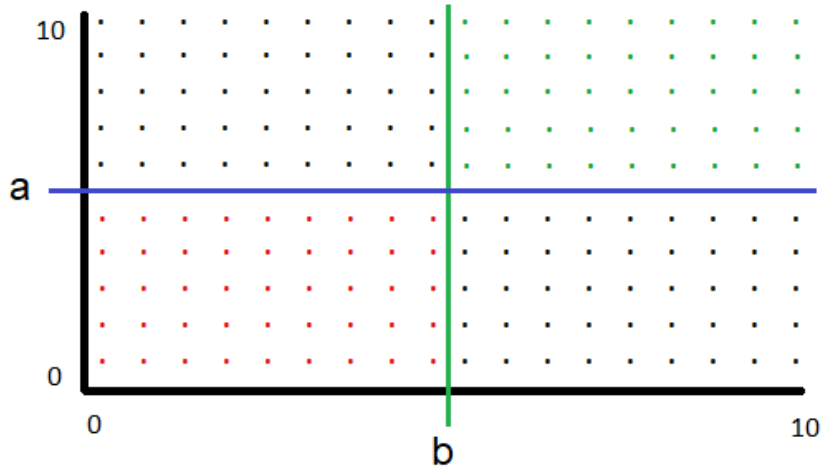



*Figure 3.15: Clustering of Individuals as the Output of the Model*

## 3.4. Validation

### 3.4.1. Evaluating Performance of ETAM

The results of statistical learning models must be evaluated for their prediction accuracy. This helps to understand the prediction power of the model when applied to new data. In order to perform the evaluation, accuracy metrics must first be defined. Evaluation should be done using data which have not be used for training. If the model is overfitted, the result of evaluation by training data will present the model as a very good one, but it will indeed perform very weakly in dealing with new data. ETAM is a model for predicting human decisions. The complex nature of human behavior makes it difficult to predict. In contrast to machine behavior prediction models such as those that predict machine failures, human behavior prediction models have low accuracy (Howarth, R. B. and Sanstad, A. H., 1995). The majority are barely better than guessing the decisions of the individuals (Legris, Ingham, and Collerette, 2003). It is good progress to improve the accuracy of current models even in tiny amounts or to make them more reliably applicable to different types of new data by removing limits and boundaries.

In statistics, many different metrics have been developed to examine the accuracy of a model by measuring the amount of prediction error. The error of a prediction model can be divided into two types, I and II. In ETAM, a Type I error is incorrectly predicting an individual as a customer. This is also known as a False Positive. A Type II error is incorrectly rejecting an individual as a customer. This is also known as a False Negative. Establishing a confusion table is suggested to evaluate the accuracy of prediction by the model. A confusion table is a clean and unambiguous way to present the prediction result of a classifier model. Table 3.3 shows the confusion table. It has four cells to show the number of observations predicted correctly and incorrectly. A positive event is acceptance of innovative technology, and a negative event is rejection of innovative technology.

As an example for the confusion table, look at Figure 3.16. Red individuals are real customers of innovative technology. Black individuals are real non-customers. Circles are clusters which individuals are predicted to be a part of. The red circle is the cluster which has been predicted as a market opportunity. A blue circle is the cluster of a non-market opportunity. Table 3.4 shows the filled confusion matrix for this clustering.

*Table 3.3: Confusion Table*

| | | Predicted | |
|---|---|---|---|
| | | **Reject** | **Accept** |
| **Actual** | **Reject** | True Positive | False Negative<br>Type II Error |
| | **Accept** | False Positive<br>Type I Error | True Negative |

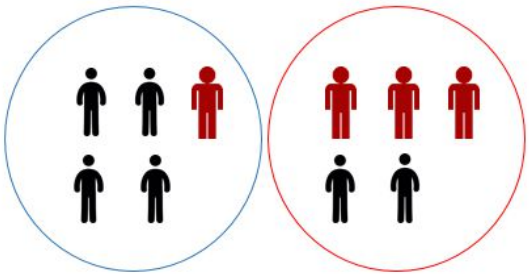


*Figure 3.16: Example of Clustering of Individuals*

*Table 3.4: Example of a Populated Confusion Table*

| | | Predicted | |
|---|---|---|---|
| | | **Reject** | **Accept** |
| **Actual** | **Reject** | 3 | 1 |
| | **Accept** | 2 | 4 |

Now that the confusion matrix has been introduced, the metrics will be discussed. The three metrics of True Positive Rate, True Negative Rate, and Balance Accuracy are suggested for evaluating the prediction accuracy. These metrics have the advantage of measuring performance for all empirical models, so they can easily be used for comparison of the ETAM prediction performance. The True Positive Rate shown in Equation 3.14, which is also known as the hit rate, measures the performance of the model at picking the right customers.

$$\text{True Positive Rate (Hit Rate)} = \frac{\sum \text{True Positive}}{\sum \text{True Positive} + \sum \text{False Negative}} \tag{3.14}$$

The True Negative Rate shown in Equation 3.15 measures the performance of the model at picking the individuals who will not accept the innovative technology.

$$\text{True Negative Rate} = \frac{\sum \text{True Negative}}{\sum \text{True Negative} + \sum \text{False Positive}} \tag{3.15}$$

Looking only at the True Positive Rate and True Negative Rate can be misleading. Generally, we expect an increase in the metric of the True Positive Rate to result in a lower True Negative Rate if we use the same type of modeling technique. For example, by giving more weight to observations of acceptance or penalizing the rejection of right customers in the model, the True Positive Rate will increase but the True Negative Rate will decrease. There are some limits, and trying to boost one can result in an overfitted model that will not predict well when fed with new data which have not been used for training. To solve this problem, use of a new set of data to calculate metrics is highly recommended. In ETAM, all evaluation metrics should be calculated using the previously discussed validation data set, which is 20% of all observations and which was not used for training.

The last introduced metric is Balanced Accuracy, defined as shown in Equation 3.16. This shows the overall performance of the model by averaging the True Positive Rate and the True Negative Rate.

$$\text{Balanced Accuracy} = \frac{\text{True Positive Rate} + \text{True Negative Rate}}{2} \quad (3.16)$$

## 3.4.2. Defining Implied Discount Rate and Payback Threshold

As mentioned in Chapter 2, many empirical prediction models use the implied discount rate to understand and predict customers. Some of them use the implied discount rate to predict customers based on the assumption that customers are Economy Rational individuals (Hausman, 1979). Others including Gilbert E. Metcalf and Kevin A. Hassett (1993) used the same concept to reject the Rational Choice Theory. This study calculates the implied discount rate of customers in order to compare performance of ETAM with empirical models.

An Economy Rational individual will accept the innovative efficiency technology if its gain is bigger than its premium cost. The gain should be calculated in a monetary scale of future savings on energy (Gilbert E. Metcalf and Kevin A. Hassett, 1993). Given the amount of an individual's yearly gain and the length of time the technology will be in use, the minimum implied discount rate for each individual can be calculated by using iterations and equation 3.17.

$$\sum_{0}^{n} \frac{y}{(1+R)^n} = \text{Premium cost of the technology} \quad (3.17)$$

$n$ Product service life

$R$ Implied discount rate

$y$ Yearly gain from cost saving in energy

The calculated implied discount rate is the minimum amount expected by a customer to consider the choice of an innovative efficiency product. Also, the payback threshold can be calculated via Equation 3.18, assuming the implied discount rate to be zero and given the amount of yearly gain.

$$\mathrm{N} = \frac{\text{Premium cost of the technology}}{\mathrm{y}} \tag{3.18}$$

$\mathrm{N}$ Payback threshold

The payback threshold illustrates the minimum expected service life of the technology required to pay off its premium cost.

## 3.4.3. Prediction Based on Rational Choice Theory

The developed ETAM database includes information regarding the amount of usage of the technology by individuals, the price of energy, and the average life of the technology. This information may be used to calculate the implied discount rate of each individual who has already accepted the new efficiency technology in the training data set. In this paper, VBA was used for coding iterations. The average of the implied discount rate can predict acceptance based on the assumption that customers are Economy Rational individuals. At the end, the confusion matrix should be developed using the validation data set to calculate performance metrics.

## 3.4.4. Prediction Based on TAM Model

The developed database holds attributes which are suggested by TAM for prediction in the Beliefs and Values category. A nominal regression can be used to predict acceptance of the technology in the training data. Independent attributes are all captured attributes that indicate an individual's viewpoint on the usefulness of the innovative technology. This viewpoint can include cost, quality, and alternatives. To compare the performance of the model, the confusion matrix should be developed, and three introduced metrics should be calculated using the validating set of data.

### 3.4.5. Comparing the Accuracy of ETAM, RC, and TAM

ETAM performance evaluation requires that all calculated metrics be compared to each other and interpreted as illustrated in Table 3.5. It may be concluded that one model is superior in all metrics or only a few. In the next chapter, a case study is performed to illustrate the power of ETAM.

*Table 3.5: Performance Comparison*

| Comparison Table | True Positive Rate | True Negative Rate | Balanced Accuracy |
|---|---|---|---|
| TAM | % | % | % |
| RC | % | % | % |
| ETAM | % | % | % |

# Chapter 4: Case Study

This chapter illustrates implementation of ETAM with a study of hybrid car sales in the state of California. Hybrid vehicles are equipped with a battery, which is charged using wasted energy from brakes, and they have an electromotor which uses the saved energy in the battery to assist the combustion engine for acceleration (U.S. Department of Energy, 2018). By the National Highway Transportation and Safety Administration's definition, a hybrid car includes any vehicle that has "an internal combustion engine and one of several possible alternate sources of propulsion" (NHTSA, 2013). However, in this research the term is used only for electric-gasoline hybrid vehicles.

## 4.1. Background

Any type of transportation, including a hybrid car, which uses renewable or regenerated energy can help to improve sustainability. The Connecticut General Assembly (CGA) measured emissions in gas engine vehicles and in their comparable hybrid vehicles. In the compact vehicle class, a reduction of 10% in emissions was recorded. This reduction in emission increased to 21% for large sport utility vehicles (SUV). The International Energy Agency (IEA) estimated that the transportation system is 95% dependent on fossil petroleum (2012). Also, transportation produces 20% of greenhouse gas emissions (IEA, 2012). These numbers prove the importance of accepting hybrid technology to preserve the earth's resources and progress in sustainability.

When looking at the other advantages of the hybrid vehicle, the fuel efficiency is a well-known selling point. A study of 2009 year model vehicles performed by the Environmental Protection Agency (EPA) found that passenger hybrid cars like the Prius or Civic can go 45% to 84% farther with a gallon of fuel than their non-hybrid counterparts, based on a driving cycle of 45% highway driving and 55% city driving. Expanding the market of hybrid cars would be progress in the sustainability which has been defined by the WCED. The

District of Colombia, California, Massachusetts, Oregon, Pennsylvania, and Washington are the six major states in the United States having sustainable transportation plans in effect (Lee et al., 2002; Jeon et al., 2007; Portney, 2002; Zhoun, 2012), and a part of their incentive policies for sustainable transportation is targeted at the sales of hybrid cars and overcoming customers' resistance to buying this innovative technology. Among these states, California achieved the highest number of hybrid car sales in 2009 with 55,553 hybrid vehicles sold. The followers in the United States were New York with 15,438 and Florida with 14,949. The state of California reached an even higher number of 91,417 hybrid car sales in 2007 (hybridCars.com, 2008-2009).

The most widely studied barrier in acceptance of hybrid technology by consumers is price. A hybrid car costs on average $5,390 more than its equivalent make and model equipped with a conventional engine (Yizao Liu, 2014). To help customers with the upfront cost of owning a hybrid car, the Internal Revenue Service (IRS) provided a $2,000 taxable income deduction to an alternative fuel vehicle purchase according to HR 1308, Section 319 of the Working Families Tax Relief Act of 2004 (Law No:108-311; Thomas, 2003). In 2005, incentives increased by the Energy Policy Act (Law No: 109-58; Barton, 2005). The Energy Policy Act established a federal income tax credit of up to $3,400 for the purchase of a new hybrid vehicle (Alan Jenn et al., 2013). Further, since December 31, 2010 electric and plug-in-hybrid vehicles are eligible for a federal income tax credit of up to $7,500 (www.fueleconomy.gov). This means that much, if not all, of the upfront cost of a hybrid can be recovered via incentives.

Unfortunately, most current policies for motivating sustainable transportation and reducing environmental impact of transportation have been rather ineffective because they have disregarded the behavioral aspects of travelers (Garcia-Sierra et al., 2015). The highest hybrid car market share till the day of the writing of this paper occurred in 2013, and this share was only 3.19% of the total year sale, which was equal to 495,534 hybrid cars (Alternative Fuels and Advanced Vehicle Data Center, 2015).

While the amount of publicly available data related to hybrid cars and green solutions are limited, the State of California Department of Motor Vehicles and California Center for Sustainable Energy provide useful information to researchers on their website. Introduced methodology in this research should be applicable to all states, but the state of California

has been selected for this case study for the previously mentioned reason. Using the same methodology and model in other states may result in a different conclusion for those states, especially when considering differences in cultural, geographical, job market, financial, and political attributes.

## 4.2. Data Collection

### 4.2.1. Incorporating the Comprehensive Set of Attributes

Buying a hybrid car is a rare event, which makes the process of data collection more challenging. The methodology introduced in this research is used to incorporate a comprehensive set of attributes from different resources. Attributes are captured from the National Household Travel Survey (NHTS), automotive manufacturer websites (Toyota, Nissan, Honda, Ford, Chevrolet, Mercury, Cadillac, BMW, Mercedes, and Hyundai), the California Center for Sustainable Energy (CSE), the State of California Department of Motor Vehicles (DMV), the U.S. Department of Energy (U.S. DOE), and the IRS. The latest set of data available from the U. S. Department of Transportation Federal Highway Administration at the time of this study is from the year 2009.

According to the breakdown of input attributes in ETAM, Figure 3.1, a total of 72 attributes were extracted from the above sources. See Table 4.1. Each captured attribute belongs to one of the categories of input attributes introduced in ETAM. In addition to these attributes, five more variables including the response variable were captured. These will be used for filtering and validating the database later in Section 4.2.3. These variables indicate the type of technology used in the engine by the owner, the type of technology used in the engine by the manufacturer, licenses plate type, state of residency, and vehicle type. For more information regarding the relation between extracted attributes and the source of information, see Appendix 1 of this study. Appendix 1 maps the attributes, their sources, and input attribute categories.

*Table 4.1: Captured Attributes*

| # | Attribute | Category of Attribute | # | Attribute | Category of Attribute |
|---|---|---|---|---|---|
| 1 | Race | Demographic | 37 | Total number of trips to school | Type of usage |
| 2 | Count of household | Demographic | 38 | Total number of trips to medical center | Type of usage |
| 3 | Severe medical condition | Demographic | 39 | Total number of trips for shopping | Type of usage |
| 4 | Primary activity | Demographic | 40 | Total number of trips for family activity | Type of usage |
| 5 | Hispanic or non-Hispanic | Demographic | 41 | Total number of trips for transporting others | Type of usage |
| 6 | Own or rent housing | Economic | 42 | Total number of trips for social activity | Type of usage |
| 7 | Total income | Economic | 43 | Total number of trips for meals | Type of usage |
| 8 | Count of vehicles | Economic | 44 | Total number of trips for other | Type of usage |
| 9 | Work status | Occupation | 45 | Total number of trips for parking at public transit | Type of usage |
| 10 | Fixed work space | Occupation | 46 | Average time at destination | Type of usage |
| 11 | Full/part time work | Occupation | 47 | Total number who used interstate | Type of usage |
| 12 | Flexible work time | Occupation | 48 | Total number who paid toll | Type of usage |
| 13 | Self employed | Occupation | 49 | Trips in a weekend | Type of usage |
| 14 | Frequency of work from home | Occupation | 50 | Total number who used public transit | Type of usage |
| 15 | Distance to work | Occupation | 51 | Day of travel | Type of usage |
| 16 | Option to work at home | Occupation | 52 | Count of trips in a week | Amount of usage |
| 17 | Minutes from home to work | Occupation | 53 | Annual miles driven | Amount of usage |
| 18 | Usual arrival time at work | Occupation | 54 | Gas price | Energy unit cost |
| 19 | Highest grade completed | Education | 55 | Workers per square mile | Population work and wealth status |
| 20 | Age of vehicle | Habits | 56 | Percent renter | Population work and wealth status |

*Table 4.1 Continued*

| # | Attribute | Category of Attribute | # | Attribute | Category of Attribute |
|---|---|---|---|---|---|
| 21 | Vehicle model year | Habits | 57 | Population per square mile | Urban/rural location |
| 22 | Number of bike trips | Habits | 58 | MSA population size for the home address | Urban/rural location |
| 23 | Number of walk trips in a week | Habits | 59 | Size of urban area in which home address is located | Urban/rural location |
| 24 | How often public transportation is used | Habits | 60 | Census division classification for home | Urban/rural location |
| 25 | Average number of people in vehicle | Habits | 61 | Census region classification for home address | Urban/rural location |
| 26 | Number of times made purchase via internet in past month | Habits | 62 | Houses per square mile | Urban/rural location |
| 27 | Number of internet purchases delivered to home | Habits | 63 | Home address in urbanized area | Urban/rural location |
| 28 | View on price of travel | Beliefs and values | 64 | Household in urban/rural area | Urban/rural location |
| 29 | View on highway congestion | Beliefs and values | 65 | Housing units per square mile | Urban/rural location |
| 30 | View on access or availability of public transit | Beliefs and values | 66 | Population per square mile | Urban/rural location |
| 31 | Most important transportation issue | Beliefs and values | 67 | MSA heavy rail status | Urban/rural location |
| 32 | View on safety concerns | Beliefs and values | 68 | Federal tax incentive | Federal |
| 33 | Frequency of internet use in past month | General knowledge through media | 69 | State tax incentive | State |
| 34 | Average number of passengers in observed trips | Type of usage | 70 | Access to HOV | State |
| 35 | Average trip distance | Type of usage | 71 | Number of available Hybrid car models | Diversity of products |
| 36 | Total number of trips to work | Type of usage | 72 | Market share of Hybrid car | Market share |

## 4.2.2. Developing the Relational Database

The primary database in this case study includes about 1,040,000 trip data records, 308,000 individual data records, 150,000 household data records, 309,000 vehicle data records, and engine type specifications of all vehicle models sold from 2002 to 2009 in the United States. To relate the information from different resources, the relational database was developed as guided by ETAM. A total of 11 tables were used to store data. As discussed in Section 3.2.2. of this study, two of the 11 tables are the result of aggregating the detailed usage information and the history of gas price in the state of California. For each individual, only those trips that the individual himself was in his car as a passenger or driver are considered to be valid trips for aggregation. See Figure 4.1. Blue triangles show the aggregated tables, and blue squares represent the rest of the tables. Data sources for each table and their foreign keys to establish relations are illustrated in Figure 4.1 as well. Running a query to get the value of all attributes for one row of observations from tables will give a row of data for an individual with a unique combination of individual identification number, household identification number, and vehicle identification number.

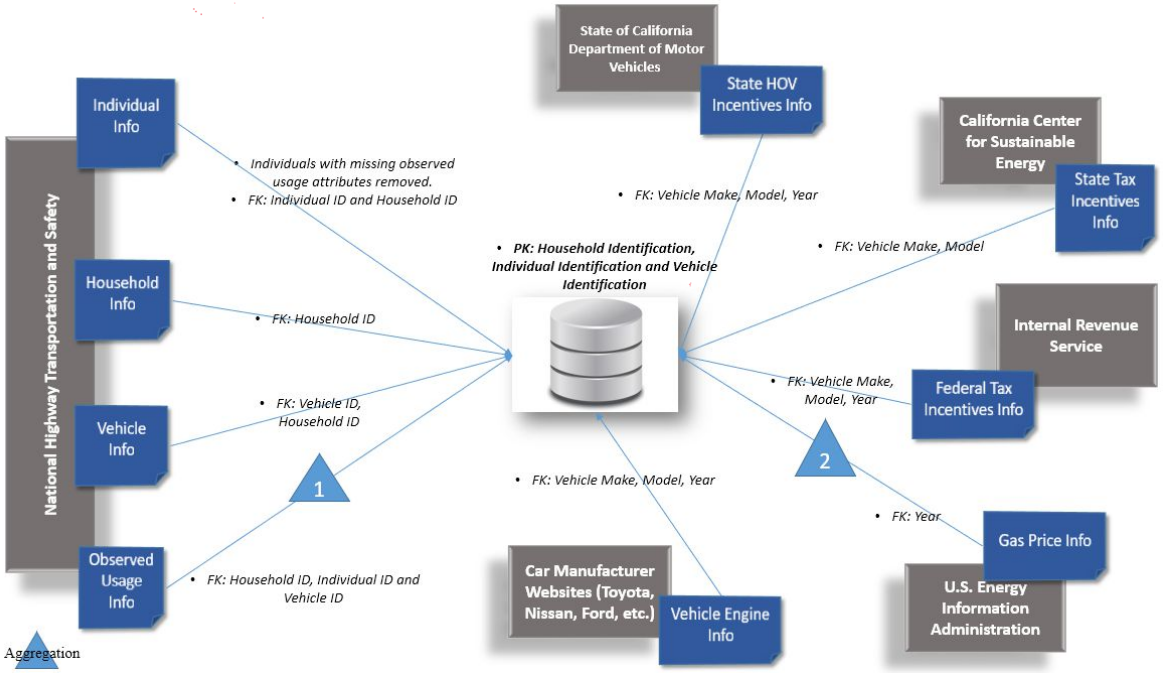



*Figure 4.1: Data Sources to Establish the Database*

## 4.2.3. Validating the Database

### 4.2.3.1 Ensuring Integrity of Data

The scope of this study includes only the state of California. as a result, information related to individuals not residing in the state of California was deleted from the database. Individuals younger than 18 years old are considered minors and need parent or guardian permission to enter a contract. Otherwise, they will not be held to their contractual obligations under law. Consequently, information related to this group of individuals was deleted from the database. Used car buyers have different priorities and motivations and are not the focus of this study. Records of information related to those who bought their vehicles used were removed from the data. Since vehicles with commercial plates are purchased by businesses and not by the individuals who use them, the records of these vehicles were removed from the database as well. Moreover, the records related to vans, trucks, golf carts, and motorcycles were dropped from the database. After defining the scope of the study and dropping unrelated information, the database was evaluated for integrity. To ensure integrity of data, individuals with no vehicle information were removed from the database. Also, individuals with missing observed usage attributes were removed from the database. When the database was cleaned, the number of usable individual records was reduced from 308,901 to 4,547. Missing data were addressed differently for continuous and categorical attributes according to Section 3.2.3.1.2. of this study. Then, vehicle information provided by owners was validated against vehicle information downloaded from vehicle manufacturers' websites. If there was a non-solvable conflict between the individual response and vehicle manufacturer data regarding the vehicle information, the individual information was removed from the database because there is no opportunity to contact them directly and resolve the conflict.

### 4.2.3.2. Removing Redundant Information

Capturing data from different sources may result in redundant attributes. Redundant information was removed in accordance with ETAM guidelines. See Table 4.2. One attribute from each pair of redundant attributes was deleted.

*Table 4.2: Redundant Information*

| Redundant to Delete | Redundant to Keep |
|---|---|
| Household in urban/rural area | Home address in urbanized area |
| Hispanic or non-Hispanic | Race |
| Vehicle model year | Vehicle age |

#### 4.2.3.3. Ensuring Database Represents the Real Word

Weights provided by NHTS, which is widely used by other researchers, was used as the primary weight to change data to represent an unbiased sample of the state of California. The auxiliary variables used by NHTS to generate weight are race, tenure, geographic area telephone exchange frame for three months, and time period of travel. The final weight was calculated based on the ETAM guideline presented in Section 3.2.3.3. using the primary weight provided by NHTS.

## 4.3. Prediction

### 4.3.1. Dividing the Data into Two Sets

Rows of data were marked randomly for training and validation use by a ratio of 4:1 according to ETAM. The number of observations in the training data set is 3,581, and 966 rows of data were dedicated for validation. The larger data set was used for training the model, and the smaller set of validation data was kept untouched for evaluating the performance of the model.

### 4.3.2. Applying the Decision Tree and Ensuring its Accuracy

The decision tree was applied using JMP software by SAS. The result of the decision tree is illustrated in Figure 4.2. The decision tree results in eight leaves. To evaluate the accuracy of the tree in accordance with ETAM, all leaves were checked for the minimum

required number of observations. The minimum required number of observations in each leaf is calculated to be 0.25% to 1% of the total number of observations in the training data set, which is roughly nine to 36 observations. The smallest leaf, which is Cluster Number 4, holds 70 observations. This is far above the required minimum.

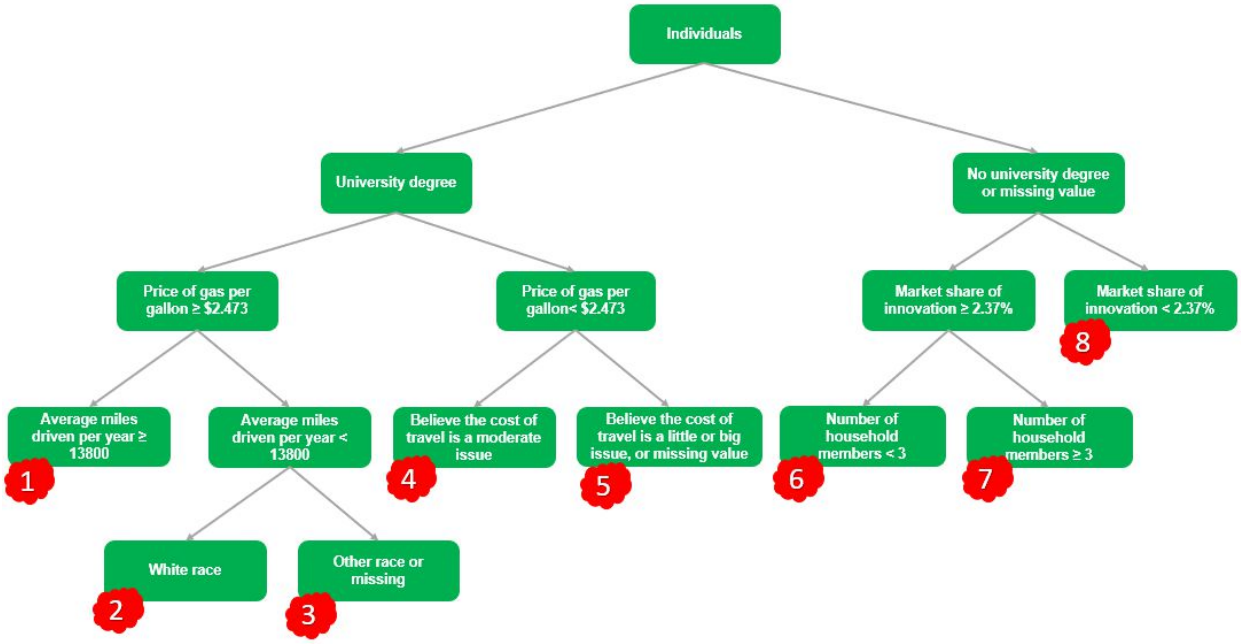



*Figure 4.2: Decision Tree for Clustering of Individuals*

The biggest leaf, which is Cluster Number 8, holds 1,149 observations. Figure 4.2 is used to answer questions regarding the characteristics of customers such as who, where, when and why. Leaf, cluster, and market segment are used interchangeably in this study.

## 4.3.3. Defining the Probability of Acceptance in a Leaf and Evaluating Clusters for Market Opportunity

The number of those who accepted the innovative efficiency technology and the calculated probability of acceptance for each leaf is shown in Table 4.3. The market share of each cluster is calculated by dividing the number of individuals who accepted the

innovative technology in each leaf by the total number of innovative technology customers.

*Table 4.3: Details of Clusters*

| Cluster | Number of Individual Accepted | Total Number of Individuals | Probability of Acceptance | 95% Confidence Interval | | Market Share of Innovative Technology |
|---|---|---|---|---|---|---|
| 1 | 125 | 427 | 29.27% | 25.15% | 33.76% | 35.01% |
| 2 | 104 | 652 | 15.95% | 13.34% | 18.96% | 29.13% |
| 3 | 10 | 134 | 7.46% | 4.10% | 13.19% | 2.80% |
| 4 | 7 | 70 | 10.00% | 4.93% | 19.23% | 1.96% |
| 5 | 33 | 646 | 5.11% | 3.66% | 7.09% | 9.24% |
| 6 | 39 | 278 | 14.03% | 10.44% | 18.60% | 10.92% |
| 7 | 11 | 225 | 4.89% | 2.75% | 8.54% | 3.08% |
| 8 | 28 | 1149 | 2.44% | 1.69% | 3.50% | 7.84% |

The prior probability of acceptance in the training set of data was calculated as instructed in Section 3.3.5. The prior probability is 9.97% among observations. Clusters 1, 2, and 6 are considered market opportunities, while the acceptance rate in Cluster 4 is roughly equal to the prior probability of acceptance. Cluster 1 has the largest market share of innovative technology and also has the highest probability of acceptance. Cluster 2 is the second largest market of innovative technology with a considerably lower probability of acceptance rate. Cluster 6 is the smallest market opportunity cluster. See Figure 4.3.

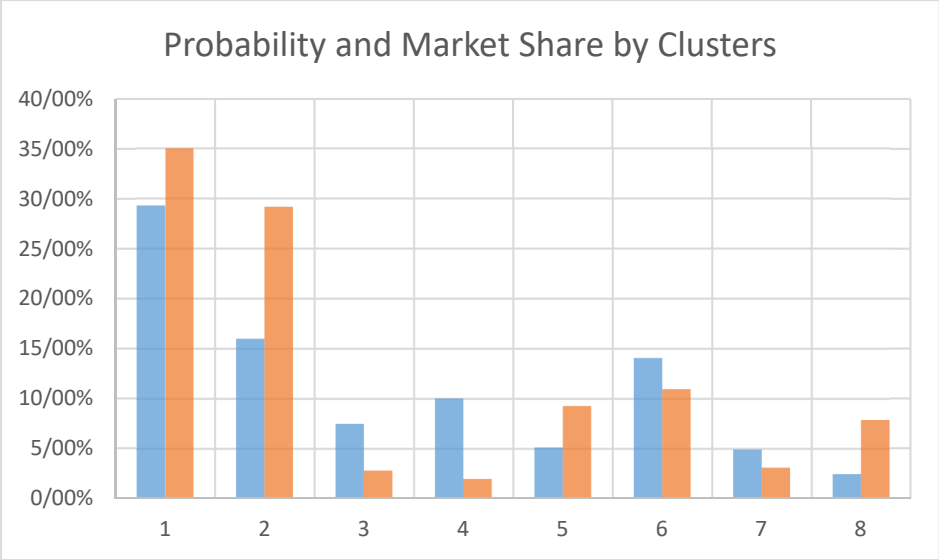



*Figure 4.3: Probability vs Market Share of Clusters*

*(Blue is the Probability of Acceptance and Orange is the Market Share)*

### 4.3.4. Answer the Questions: Who will Accept and When, Where, and How?

As can be seen in Figure 4.2 and Table 4.3, the amount of education is the most important factor in accepting the new efficiency technology. 65% of the market share of new efficiency technology is driven by individuals in Clusters 1 and 2 who have a university degree.

Individuals in Cluster 1, which is the biggest cluster at 35% and which has the highest probability of acceptance at 29%, not only are educated, but also consider the price of gas and their annual miles driven to make an economical decision. The acceptance probability in this cluster is roughly 3 times the average probability of acceptance among observations of this study. These individuals are willing to accept the technology if the price of gas is higher than $2.47 per gallon and if they drive their car more than 13,800 miles per year.

Individuals in Cluster 2, which is still an important market share of innovative technology at 29%, are only sensitive to the price of energy. Their probability of acceptance is 16%, which is roughly 60% higher than the average probability of acceptance in observed

individuals. They will choose the innovative efficiency technology if the price of gas is equal to or higher than $2.47 per gallon. One interesting significant attribute in this cluster is race. The race of individuals in this cluster is white. These customers invest in the innovative technology while their investment in efficiency may or may not be paid off by their amount of usage.

Cluster 6 consists of 11% of the market of new technology customers. Their probability of acceptance is 14%, which is 40% higher than the average acceptance probability. These individuals do not hold a degree from a university, and they will only accept the hybrid cars if their market share is higher than 2.37% of the automotive market. This has already been shown in the Diffusion of Innovations Theory by Rogers (1962). According to the finding of this study, this theory is more useful to understand late acceptance of lower educated individuals in Cluster 6. Another interesting significant attribute in Cluster 6 is the number of household members. This cluster of customers are households with equal to or less than 2 members.

The probability of accepting the new efficient technology by lower educated individuals when the market share is lower than 2.37% is as low as 2.44%, as can be seen in Cluster 8.

The Probability of acceptance in Cluster 4 is barely higher than the average acceptance probability in observed individuals, 10% against 9.97%. This cluster is only 1.96% of the market of the efficiency technology. The significant attribute which differentiates this cluster from other clusters is how individuals in this cluster think or believe regarding the cost of travel.

### 4.3.5. Evaluate the Result for Other Information and Trends

The result of the decision tree shows that previous theories, including the Rational Choice Theory in economics, the Diffusion of Innovations Theory, and TAM, are valid only for a group of individuals. For example, Cluster 1 is a good example of Economy Rational customers, while Clusters 2 and 4 are good samples of customers who choose the efficiency technology because of their belief. Clusters 6, 7, and 8 show the effect of market share on acceptance as indicated by Rogers (1962) while other educated individuals are

not affected by the market share. The diffusion of Innovations Theory can easily be seen, as the model predicted that the lower educated individuals will accept the innovation later in time when the market share is higher than 2.37%.

Table 4.4 shows the calculated payback threshold for individuals who bought hybrid cars in each cluster. For this calculation, the average miles driven per year by each individual, the average price of gas, and the available monetary incentives at the time of purchase were considered. Monetary incentives, such as available tax credits, help to reduce the cost of initial investment and affect the payback threshold.

While there is no evidence from the result of ETAM to prove or reject that individuals in clusters calculate and consider payback threshold as a base for decision making, these thresholds are calculated to better understand characteristics of individuals in each cluster. Customers in Cluster 1, which is the only cluster differentiated by the two cuts of gas price and miles driven in the increasing directions, has a very low payback threshold of two years. This means that their investment will be paid off in two years. Customers in Clusters 2 and 6, which both are considered market opportunity, have a much higher payback threshold of five and six years for their investment.

Customers in Cluster 2 are differentiated with gas price cut in the increasing direction, which may represent a simple heuristic decision-making process with hopes to result in a better financial outcome. The payback threshold of customers in this cluster is five years, which is very close to the 4.3 years average length of vehicle ownership for new car buyers in the United States (IHS, 2006).

While Clusters 5, 7, and 8 have the lowest probability of acceptance, of these three only the two Clusters 5 and 8 have high payback threshold based on their usage. Customers in Cluster 7 has a very low payback threshold of 2.6 years.

*Table 4.4: Payback Threshold of Individuals in Clusters*

| **Cluster** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| Payback Threshold | 1.87 | 4.89 | 3.19 | 5.46 | 7.53 | 5.99 | 2.55 | 6.08 |

# 4.4. Validation

## 4.4.1. Evaluating Performance of ETAM

To evaluate the performance of ETAM, it was applied to the validation set of data. The confusion matrix which shows the number and percentage of observations predicted correctly can be seen in Table 4.5. Accuracy metrics were calculated as indicated in Section 3.4.1. The True Positive Rate for ETAM is 63.64%, which means the model predicted close to 64% of customers of hybrid cars correctly. The True Negative Rate for ETAM is 66.4% which indicates that 66% of those who reject the efficiency technology were predicted by the model correctly. The model resulted in a balance accuracy of 65.02% which means that the ETAM predicted acceptance and rejection of the efficiency technology by an accuracy of 65%. In other words, 65% of individual decisions are predicted correctly.

*Table 4.5: Performance of ETAM*

| ETAM | | Predicted | |
|---|---|---|---|
| | | Reject | Accept |
| Actual | Reject | 583 | 295 |
| | Accept | 32 | 56 |

| ETAM | | Predicted | |
|---|---|---|---|
| | | Reject | Accept |
| Actual | Reject | 66.40% | 33.60% |
| | Accept | 36.36% | 63.64% |

## 4.4.2. Prediction Based on Rational Choice Theory

Prediction via the Rational Choice Theory requires calculation of the implied discount rate by customers. According to the Institute for Highway Safety (IHS, 2006), 4.3 years is the average time of ownership of new vehicle buyers. In addition, a hybrid car on average costs $5,390 more than its equivalent make and model equipped with a conventional engine (Yizao Liu, 2014).

The median of implied discount rate by customers of innovative efficiency technology in the training set of data was calculated as 31.5% by plugging the values for average years

of new vehicle ownership and premium price of innovative technology into Formula 3.18. The result from predicting acceptance of efficiency technology in the validation set of data using the calculated implied discount rate is shown in Table 4.6.

The accuracy metrics were calculated as indicated in Section 3.4.1. The True Positive Rate for the model based on the Rational Choice Theory is 44.32%, which means the model predicted only 44% of customers of hybrid car correctly. The True Negative Rate for ETAM is 79.73% which indicates that close to 80% of those who reject the efficiency technology were predicted by ETAM correctly. This model results in a balance accuracy of 62.02% which means that ETAM predicted acceptance and rejection of the efficiency technology with an accuracy of 62%.

*Table 4.6: Performance of the Rational Choice Theory Model*

| RC | | Predicted | |
|---|---|---|---|
| | | Reject | Accept |
| Actual | Reject | 700 | 178 |
| | Accept | 49 | 39 |

| RC | | Predicted | |
|---|---|---|---|
| | | Reject | Accept |
| Actual | Reject | 79.73% | 20.27% |
| | Accept | 55.68% | 44.32% |

This high accuracy is driven by the power of the model to predict rejection of efficiency technology, not acceptance of it.

## 4.4.3. Prediction Based on TAM

To implement TAM, beliefs and values attributes were used as the input of a nominal regression model to predict acceptance. Then, the model was applied to predict acceptance using the validation set of data. The result is presented in Table 4.7.

The True Positive Rate for TAM is 60.23% which means the model predicted close to 60% of customers of hybrid car correctly. The True Negative Rate for TAM is 41.91% which indicates that close to 42% of those who reject the efficiency technology were

predicted by the model correctly. The model resulted in a balance accuracy of 51.07% which means that the TAM model predicted acceptance and rejection of the efficiency technology with an accuracy of 51%. While the balance accuracy of the TAM model is low, making this model poor, its accuracy in predicting acceptance is respectable.

*Table 4.7: Performance of TAM*

| TAM | | Predicted | |
|---|---|---|---|
| | | Reject | Accept |
| Actual | Reject | 368 | 510 |
| | Accept | 35 | 53 |

| TAM | | Predicted | |
|---|---|---|---|
| | | Reject | Accept |
| Actual | Reject | 41.91% | 58.09% |
| | Accept | 39.77% | 60.23% |

## 4.5. Sensitivity Analysis

This section evaluates the proposed decision tree from three different perspectives. The first evaluates the model's sensitivity to the chosen method for handling missing data. The second evaluates the model's sensitivity to the different values of the minimum allowed node observations and the stopping rule. The third studies the decision tree's sensitivity to using a selective attribute for the first split.

### 4.5.1. Sensitivity to Missing Values Handling Technique

The goal of this section is to evaluate the sensitivity of the model to choosing other techniques for handling missing values such as deleting the records with missing values or imputing the missing values.

As discussed in Section 3.2.3.1.2., of this this study, handling of missing values can be done using two different techniques. First, it can be done by dropping the records with missing data. Second, it can be done by imputing missing values and replacing them.

ETAM considers missing data as informative missing information and proposes a technique to handle them. For categorical attributes, missing values are introduced as a new level in each attribute. For continuous attributes, the observations and records are sorted ascending according to the values of the attribute which has missing values. Two new attributes are generated by adding the missing values; once at the top of the sorted values and once at the bottom. See Figure 3.8 for more information. The output of ETAM using the proposed technique for handling missing data was shown earlier in Figure 4.2 and Table 4.5.

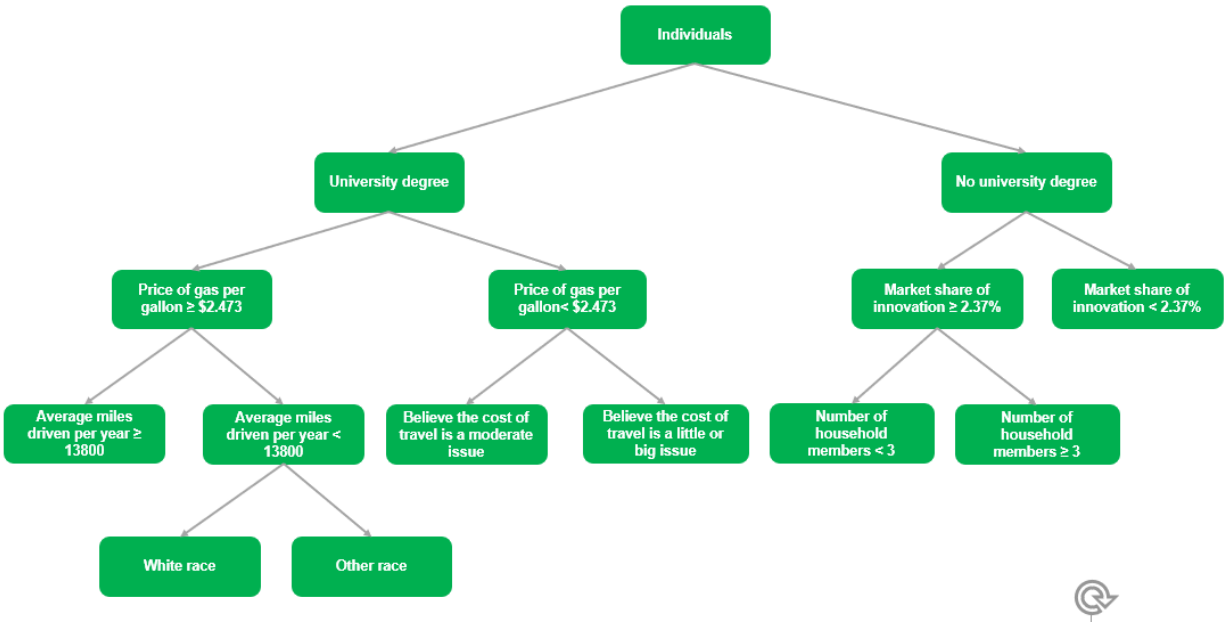



*Figure 4.4: Decision Tree Using Imputed Attributes*

The number of data cells with missing information is estimated as 10% of the total number captured in the case study of this paper. Each observation has at least two missing values. As a result, it is not practical to drop observations with missing values. Each data cell is the observed value of an attribute for an individual.

Instead of deleting observations with missing cell values, the missing values are imputed by replacing them with the median of values in each attribute. The established decision tree based on imputed attributes looks very similar to the one from the original run. The

only difference is that there is no level known as “missing” in splits since the “missing” level no longer exists as a category of attributes. See Figure 4.4.

Table 4.8 shows the calculated confusion table for the decision tree which uses imputed attributes as its input. Table 4.9 shows the performance comparison of the decision tree that uses imputed attributes and the original run that considers missing data as informative information. Changing the method of handling missing data has not changed the outcome of the model significantly.

*Table 4.8: Confusion Table for Imputed Decision Tree*

| | | Actual | |
|---|---|---|---|
| | | Reject | Accept |
| Predicted | Reject | 580 | 298 |
| | Accept | 32 | 56 |

## 4.5.2. Sensitivity to Stopping Rules

As discussed earlier in Section 3.3.2.2, the decision tree in ETAM uses two stopping rules. One looks to see if the amount of correct prediction improves in the next 10 splits, and the other looks for the minimum number of observations in leaves to help prevent the problem of an over fitted model.

*Table 4.9: Performance Comparison of Missing Data Handling Techniques*

| Performance Metrics | True Positive Rate | True Negative Rate | Balanced Accuracy |
|---|---|---|---|
| **Original ETAM** | 63.64% | 66.40% | 65.02% |
| **ETAM-Imputed (Median)** | 63.64% | 66.06% | 64.85% |

Berry and Linoff (1999) suggest 0.25% to 1% percent of observations as the minimum number of observations in a leaf. In this section different values within the suggested range by Berry and Linoff (1999) are examined to better understand the sensitivity of the model.

Since the original run of the case study model was stopped with the other stopping rule and not the minimum number of observations in a leaf, this rule is relaxed to be able to check the effect of the minimum number of observations in a leaf. The model continues splitting just till it reaches the minimum number of observations in a leaf which has been set. The performance comparison of the runs, in addition to the number of splits occurring in each run before reaching the minimum number of observations, is shown in Table 4.10. The error term in Table 4.10 is the number of times the model predicts incorrectly using the validation data set. In case of having a binary response, the sum of the difference between the predicted values and the actual values, the sum of squared error, is equal to the number of incorrect predictions.

*Table 4.10: Performance Comparison of Different Stopping Rules*

| Minimum Number of Observations in a Leaf | Error | Number of Splits | True Positive Rate | True Negative Rate | Balanced Accuracy |
|---|---|---|---|---|---|
| **Original** | 327 | 7 | 63.64% | 66.40% | 65.02% |
| **1.000%** | 329 | 12 | 63.64% | 66.17% | 64.90% |
| **0.750%** | 329 | 12 | 63.64% | 66.17% | 64.90% |
| **0.500%** | 329 | 12 | 63.64% | 66.17% | 64.90% |
| **0.375%** | 370 | 16 | 68.18% | 61.05% | 64.61% |
| **0.250%** | 373 | 30 | 68.18% | 60.71% | 64.44% |

As can be seen in Table 4.10, reducing the minimum of observations in leaves does not help to increase the prediction accuracy. It indeed makes the model more complex. A model with more splits is considered a more complex model. Reducing the minimum of observations also increases the model's number of prediction errors when using new data to predict the acceptance of the innovation. This outcome is expected since more complex models have a higher tendency to make prediction errors when facing new data. The only advantage of relaxing one of the stopping rules and reducing the minimum number of observations in a leaf is a slightly higher True Positive Rate. However, this has been

achieved at the price a much more complex model and a higher number of incorrect predictions. A simpler model with fewer splits is preferred. Thus, if the amount of improvement with more splits is not significant, it is strongly suggested to stick with fewer splits.

### 4.5.3. Sensitivity to Selective First Split

As discussed earlier in Section 3.3.2.2., the proposed model chooses the best attributes for splitting to reduce the impurity of observations in nodes. The candidate for the first split achieves the highest amount of reduction in impurity among input attributes of the model.

*Table 4.11: Performance Comparison of Selective vs Nonselective First Split*

| Performance Metrics | True Positive Rate | True Negative Rate | Balanced Accuracy |
|---|---|---|---|
| **Original ETAM** | 63.64% | 66.40% | 65.02% |
| **First split: Perceived Cost of Transportation (Attribute was selected originally as a significant one)** | 54.55% | 72.67% | 63.61% |
| **First split: Flexible Work Time (Attribute was NOT selected originally as a significant one)** | 56.82% | 68.11% | 62.46% |

To see the proposed model's sensitivity to the candidate attribute for the first split, two alternative attributes are chosen for the first split instead of the one selected by the model. The first one is selected from the attributes which have already been chosen as significant ones in the original run by the model. The other one is selected from the input attributes which have not been chosen as significant by the proposed model in the original run.
The model will split, as it is intended, after the first selective split. As can be seen in Figure 4.5 and Figure 4.6 the model tries to compensate the selection of the first attribute by selecting the best possible attributes for the next splits. Many attributes and splits that were seen in the original run, shown in Figure 4.2, can be seen in these two Figures as

well. As can be seen in Table 4.11, the performance of the new runs that include a selective split are not as good as the performance of the original run.

The new trained models have less predictive power in comparison to the proposed original one. Starting at the second split, the decision tree minimizes the node impurity in the same way as the original run to improve the prediction power.

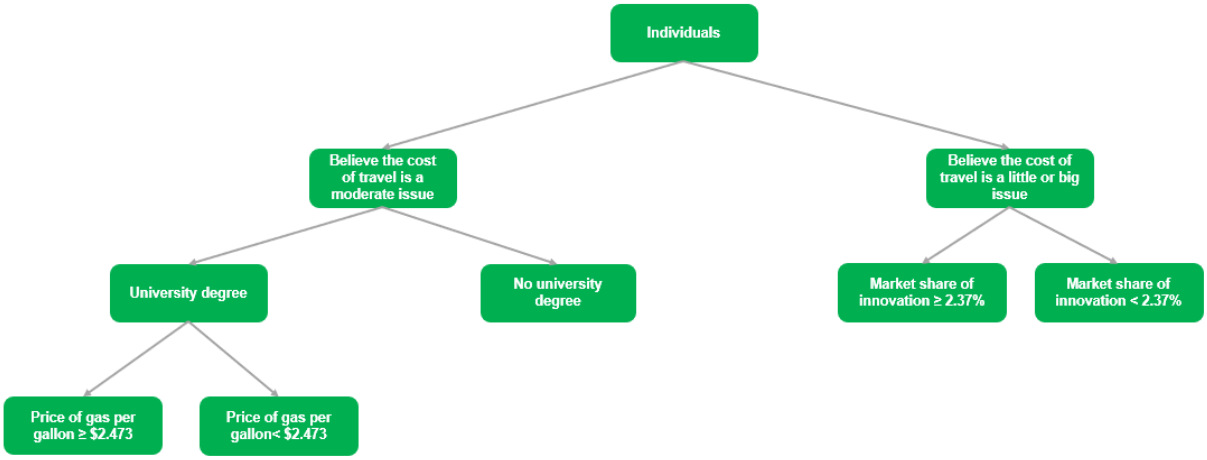



*Figure 4.5: Decision Tree with Selective First Split (Perceived Cost of Transportation)*

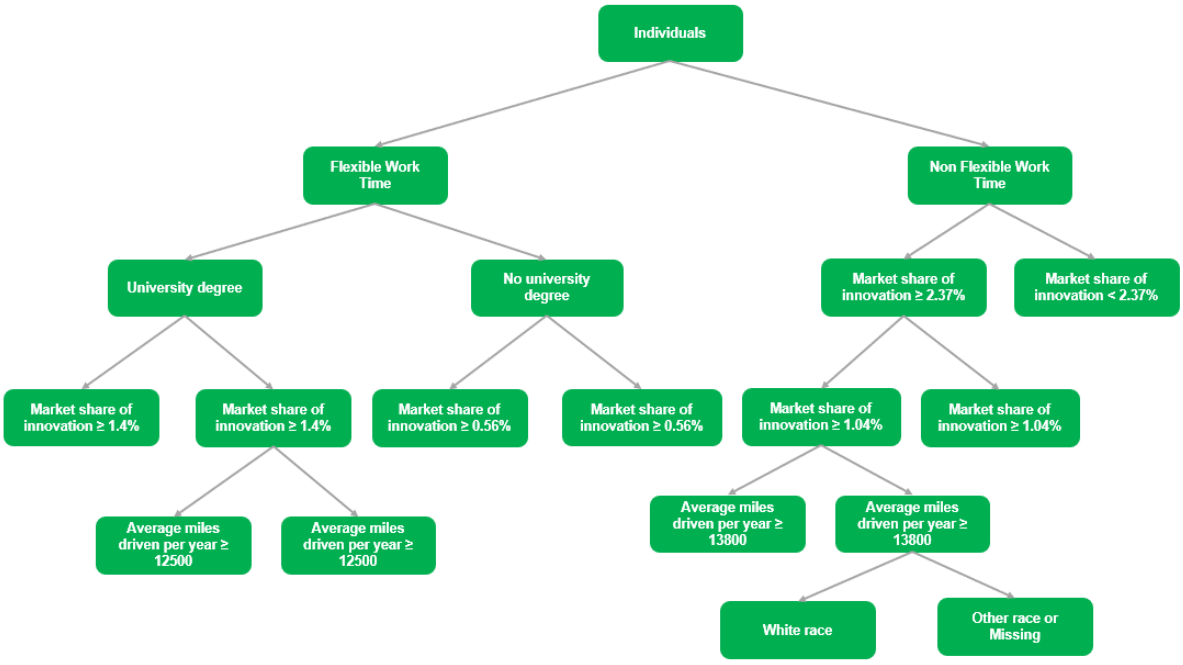



*Figure 4.6: Decision Tree with Selective First Split (Flexible Work Time)*

# Chapter 5: Conclusion

Customer resistance against new innovative technologies has been studied in many previous publications to improve the prediction power of models (Howarth, R. B. and A. H. Sanstad, 1995). The famous Diffusion of Innovations Theory by Rogers (1962) is widely used to predict market share of an innovative technology over time. It considers the importance of communication and social norms. This model does not directly predict the acceptance of the technology by an individual. The new Theory of Consumer Demand by Lancaster (1966) and the Random Utility Theory by McFadden (1976) make use of the Rational Choice Theory to predict acceptance of efficiency technology by an individual. These models failed to predict acceptance of new technology accurately. This is known as the result of non-informed and non Economy Rational customers.

Fred D. Davis introduced TAM in 1989. He used the perceived views and beliefs of individuals to predict if they would accept innovative technology. As Legris, Ingham, and Collerette (2003) highlighted, TAM only accounts for 40% of the usage of innovative technology.

This study introduced a new modeling technique named as ETAM. ETAM progresses empirical models by considering heterogeneity of customers and by relaxing many of their assumptions such as the Rational Choice Theory. ETAM is the first model to consider a comprehensive set of input attributes. ETAM is capable of simulating decision processes of customers. Table 5.1 compares the performance of ETAM, TAM, and RC. The accuracy metrics are calculated as indicated in Section 3.4.1.

ETAM achieves the highest balanced accuracy, 65%, which is an indication of how accurate it is in predicting acceptance and rejection of the innovative efficiency technology. ETAM also achieves the highest True Positive Rate, 64%.

The model based on the Rational Choice theory is the next best model considering the balanced accuracy of the model, 62%. However, this model has a very low True Positive Rate. The True Positive Rate of 44% means that this model is less accurate than just guessing by chance who will accept the technology. Surprisingly, this model works well

to predict who will not accept the efficiency technology. This model achieves 79.7% for the True Negative Rate, which helps to achieve the next highest balanced accuracy. This model works best to predict who will not accept the efficiency technology.

*Table 5.8: Performance Comparison of ETAM, TAM and RC*

| Comparison Table | True Positive Rate | True Negative Rate | Balanced Accuracy |
|---|---|---|---|
| ETAM | 63.64% | 66.40% | 65.02% |
| TAM | 60.23% | 41.91% | 51.07% |
| RC | 44.32% | 79.73% | 62.02% |

TAM achieves the lowest balanced accuracy because of its low performance in predicting those who will not accept the efficiency technology. This model achieves the next highest True Positive Rate after ETAM. TAM is a poor model since it will have many false positives in predicting acceptance of efficiency technology compared to other models. See Table 4.7.

ETAM is the best model among the three models, since it predicts acceptance of new technology with the lowest number of false positives and with a high accuracy of 65%.

The outcome of the decision tree in Figure 4.2 indicates that previous theories should be considered only for a group of individuals and not for all. The result from ETAM proves the existence of Diffusion of Innovations as theorized by Rogers (1962) for lower educated individuals. Also, ETAM shows that the perceived view of individuals is not the best attribute for predicting acceptance of the innovative efficiency technologies. 35% of customers consider price and amount of usage in choosing the efficiency technologies which supports econometrics models and proves the Rational Choice Theory (RC), at least for a large group of customers. Meanwhile, these customers have a very low payback threshold.

In marketing, the goal is to decrease the cost of advertisement by targeting the advertisement on the right cluster of individuals and increasing the acceptance rate. This helps to reduce the amount of resources and increase the efficiency of advertisement

campaigns. ETAM helps to establish a lean marketing campaign. Lean is a term from manufacturing, and it is defined as the use of different techniques to reduce waste and increase efficiency. Table 5.2 shows the number of individuals predicted by each model to be candidate customers. This table also includes the number of individuals from candidates who really accepted the innovative technology. The values in the last column are calculated by dividing the number of actual customers by the total number of customers predicted by each model. This rate shows the success rate of a campaign when using any of these models to predict customers. RC not only achieves the lowest success rate among all three models, but also results in a lower number of customers compared to ETAM. TAM beats ETAM regarding success rate (only by 1%), but ETAM results in a considerably higher number of customers if chosen by the campaign as the prediction model.

*Table 5.9: Comparison of Acceptance Rate of ETAM, TAM, and RC*

| Comparison Table | Individuals Accepted | Predicted Customers | Acceptance Rate of Targeted Individuals |
|---|---|---|---|
| ETAM | 56 | 351 | 16% |
| TAM | 39 | 217 | 17% |
| RC | 53 | 563 | 9% |

The total net profit in accepting an innovative efficiency technology by customers depends on the profit from selling each unit, the number of sales, the cost of advertisement for each individual, and number of targeted individuals for advertisement. Equation 5.1 shows how the total profit is calculated.

$$\mathrm{Max}\,(\mathrm{P}) = \mathrm{I} \times \mathrm{S} - \mathrm{C} \times \mathrm{N} \tag{5.1}$$

$\mathrm{C}$ Cost of advertisement for each individual

I Profit from selling each unit of product

N Number of targeted individuals for advertisement

P Total profit

S Number of sales

A higher number of sales and a lower number of targeted individuals for advertisement should result in increased total profit.

Depending on the unit profit from acceptance of innovative technology and cost of advertisement for each individual in the target market, ETAM, RC, or no model may be chosen to achieve the highest amount of profit. If the advertisement cost for each individual of the target market is negligible and close to zero, no model is needed. If the advertisement cost for each individual of the target market is low compared to the profit from the acceptance of the innovative technology, ETAM should be selected as a superior model. Assuming 1,000 units of currency for the profit resulting from acceptance of the innovative technology, changes in the cost of advertisement from zero to 150 units of the currency result in a different total profit if the ETAM or RC model is chosen. See Figure 5.1.

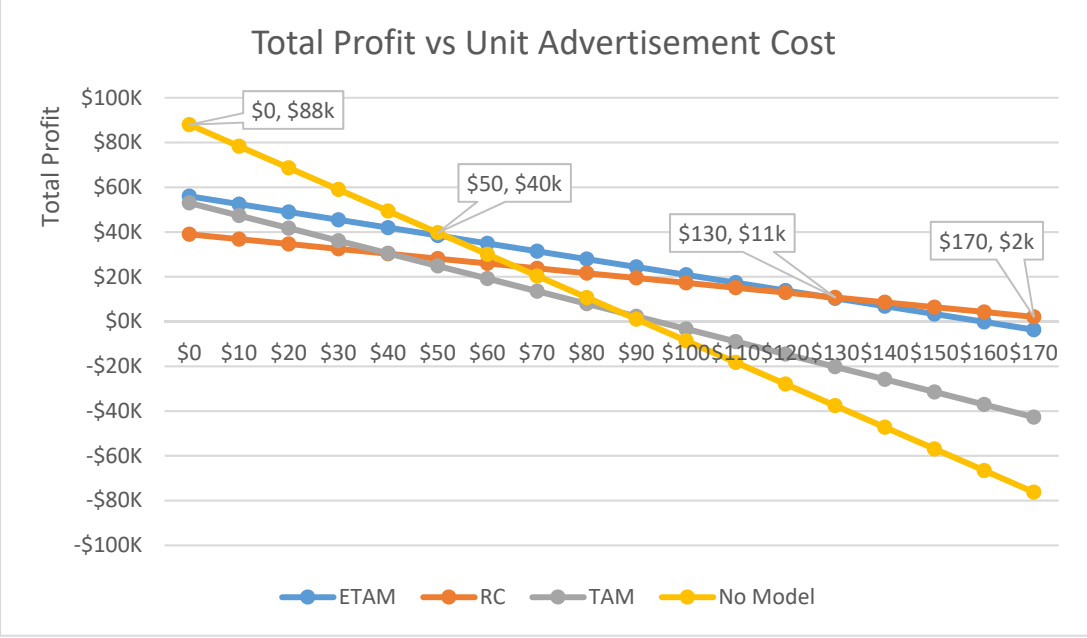



*Figure 5.4: Total Profit vs Advertisement Cost Comparison of ETAM and RC*

In Figure 5.1, the horizontal axis shows the cost of advertisement for each individual. When the advertisement cost is lower than roughly $50, 5% of profit, no model should be used. In other words, the advertisement should be done for all individuals. When the advertisement cost is lower than roughly $130 per individual, 13% of profit, but higher than $50, 5% of profit, ETAM (show in blue) results in higher profit. When the advertisement cost is higher than roughly $130 per individual,13% of profit, but lower than $50, 5% of profit, RC results in higher profit. In reality, the advertisement cost of most businesses is closer to the range of 7% to 12%, which makes ETAM the better option in most cases. If increasing the number of individuals who accept the innovative efficiency technology is the priority to the cost, then ETAM is the clear winner among the models. ETAM not only predicts acceptance very well, but also gives information regarding who are the customers, where they are located, when they will accept the new technology, why they would accept it, and what are their motivations. This would result in better and more effective use of advertising resources and communication with customers.

# References

1. Allcott, H. and M. Greenstone (2012). "Is there an Energy Efficiency Gap?" The Journal of Economic Perspectives 26(1): 3-28.
2. Arrow, K. (1962). "The Economic Implications of Learning by Doing. Review of Economic Studies, June."
3. Barnes, A. J., et al. (2016). "Promising Approaches From Behavioral Economics to Improve Patient Lung Cancer Screening Decisions." Journal of the American College of Radiology 13(12, Part B): 1566-1570.
4. Bento, A. M., et al. (2005). "The effects of urban spatial structure on travel demand in the United States." Review of Economics and Statistics 87(3): 466-478.
5. Bento, A. M., et al. (2012). "Is there an energy paradox in fuel economy? A note on the role of consumer heterogeneity and sorting bias." Economics Letters 115(1): 44-48.
6. Berry, L. (1984). "The role of financial incentives in utility-sponsored residential conservation programs: a review of customer surveys." Evaluation and Program Planning 7(2): 131-141.
7. Berry, M. and G. Linoff (1999). Mastering Data Mining: The Art and Science of Customer Relationship Management, John Wiley \& Sons, Inc.
8. Blumstein, C., et al. (1980). "Overcoming social and institutional barriers to energy conservation." Energy 5(4): 355-371.
9. Brownstone, D. and T. F. Golob (2009). "The impact of residential density on vehicle usage and energy consumption." Journal of Urban Economics 65(1): 91-98.
10. Carpenter, E. H. and S. T. Chester (1984). "Are federal energy tax credits effective? A Western United States survey." The Energy Journal 5(2): 139-149.
11. Choi, H. and I. Oh (2010). "Analysis of product efficiency of hybrid vehicles and promotion policies." Energy Policy 38(5): 2262-2271.
12. Clark, J. K., et al. (2009). "Spatial characteristics of exurban settlement pattern in the United States." Landscape and Urban Planning 90(3-4): 178-188.
13. Crawford, C. M. (2008). New products management, Tata McGraw-Hill Education.
14. Davis, F. D. (1989). "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." MIS Quarterly 13(3): 319-340.

15. Gallego, F., et al. (2013). "The effect of transport policies on car use: A bundling model with applications." Energy Economics 40: S85-S97.
16. Garcia-Sierra, M., et al. (2015). "Behavioural economics, travel behaviour and environmental-transport policy." Transportation Research Part D-Transport and Environment 41: 288-305.
17. Gillingham, K., et al. (2006). "Energy efficiency policies: a retrospective examination." Annu. Rev. Environ. Resour. 31: 161-192.
18. Gillingham, K., et al. (2009). "Energy efficiency economics and policy." Annu. Rev. Resour. Econ. 1(1): 597-620.
19. Gillingham, K. and K. Palmer (2014). "Bridging the Energy Efficiency Gap: Policy Insights from Economic Theory and Empirical Evidence." Review of Environmental Economics and Policy 8(1): 18-38.
20. Gilmore, E. A. and L. B. Lave (2013). "Comparing resale prices and total cost of ownership for gasoline, hybrid and diesel passenger cars and trucks." Transport Policy 27: 200-208.
21. Greene, D. L. (2010). How consumers value fuel economy: A literature review.
22. Greene, D. L., et al. (2014). "Analyzing the transition to electric drive vehicles in the U.S." Futures 58: 34-52.
23. Hassett, K. A. and G. E. Metcalf (1993). "Energy conservation investment: Do consumers discount the future correctly?" Energy Policy 21(6): 710-716.
24. Hassett, K. A. and G. E. Metcalf (1995). "Energy tax credits and residential conservation investment: Evidence from panel data." Journal of Public Economics 57(2): 201-217.
25. Hassett, K. A. and G. E. Metcalf (1999). "Investment with Uncertain Tax Policy: Does Random Tax Policy Discourage Investment?" The Economic Journal 109(457): 372-393.
26. Hausman, J. A. (1979). "Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables." The Bell Journal of Economics 10(1): 33-54.
27. Helfand, G. and A. Wolverton (2011). "Evaluating the Consumer Response to Fuel Economy: A Review of the Literature." International Review of Environmental and Resource Economics 5(2): 103-146.

28. Hendler, R. (1975). "Lancaster's New Approach to Consumer Demand and Its Limitations." The American Economic Review 65(1): 194-199.
29. Howarth, R. and A. H. Sanstad (1995). "DISCOUNT RATES AND ENERGY EFFICIENCY." Contemporary Economic Policy 13(3): 101-109.
30. Howarth, R. B. and A. H. Sanstad (1995). "Discount rates and energy efficiency." Contemporary Economic Policy 13(3): 101-109.
31. IEA, C. (2012). "emissions from Fuel Combustion." International Energy Agency.
32. Jaffe, A. B., et al. (2004). "Economics of energy efficiency." Encyclopedia of energy 2: 79-90.
33. Jaffe, A. B. and R. N. Stavins (1994). "The energy-efficiency gap What does it mean?" Energy Policy 22(10): 804-810.
34. Jenn, A., et al. (2013). "The impact of federal incentives on the adoption of hybrid electric vehicles in the United States." Energy Economics 40: 936-942.
35. Jeon, C. M. and A. Amekudzi (2005). "Addressing Sustainability in Transportation Systems: Definitions, Indicators, and Metrics." Journal of Infrastructure Systems 11(1): 31-50.
36. Jeon, C. M., et al. (2007). Evaluating transportation system sustainability: Atlanta metropolitan region. Proceedings of the 2007 Annual Meeting of the Transportation Research Board–CDROM.
37. Kahneman, D. (2011). Thinking, fast and slow, Macmillan.
38. Kahneman, D. and A. Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk." Econometrica 47(2): 263-291.
39. Kalton, G. (1983). "Compensating for missing survey data."
40. Kalton, G. (1983). Introduction to survey sampling, Sage.
41. Kempton, W., et al. (1992). ""I always turn it on super": user decisions about when and how to operate room air conditioners." Energy and Buildings 18(3): 177-191.
42. Kempton, W. and L. Montgomery (1982). "Folk quantification of energy." Energy 7(10): 817-827.
43. Kim, J. and D. Brownstone (2013). "The impact of residential density on vehicle usage and fuel consumption: Evidence from national samples." Energy Economics 40: 196-206.

44. Klapper, D., et al. (2005). "Another look at loss aversion in brand choice data: Can we characterize the loss averse consumer?" International Journal of Research in Marketing 22(3): 239-254.
45. Lancaster, K. J. (1966). "A New Approach to Consumer Theory." Journal of Political Economy 74(2): 132-157.
46. Lee, R. W., et al. (2002). The California general plan process and sustainable transportation planning. San Jose, CA (Mineta Transportation Institute, College of Business--BT550, San José State University, San Jose 95192-0219), Mineta Transportation Institute, College of Business, San José State University.
47. Lee, R. W., et al. (2001). "The California General Plan Process and Sustainable Transportation Planning." Planning 233: 233.
48. Legris, P., et al. (2003). "Why do people use information technology? A critical review of the technology acceptance model." Information & management 40(3): 191-204.
49. Liu, Y. (2014). "Household demand and willingness to pay for hybrid vehicles." Energy Economics 44: 191-197.
50. McFadden, D. L. (1976). Quantal choice analaysis: A survey. Annals of Economic and Social Measurement, Volume 5, number 4, NBER: 363-390.
51. Metcalf, G. E. (1994). "Economics and rational conservation policy." Energy Policy 22(10): 819-825.
52. Page, R. M. and G. E. Cole (1985). "Fishbein's Model of Behavioral Intentions: A Framework for Health Education Research and Curriculum Development." International Quarterly of Community Health Education 5(4): 321-328.
53. Portney, K. E. (2002). "Taking sustainable cities seriously: A comparative analysis of twenty-four US cities." Local Environment 7(4): 363-380.
54. Portney, K. E. (2003). "Taking Sustainable Cities Seriously-Economic Development, the Environment, and Quality of Life in American Cities." Taking Sustainable Cities Seriously-Economic Development, the Environment, and Quality of Life in American Cities: 1-284.
55. Potter, A., et al. (2002). "Investigation of the persistence of new building commissioning." Lawrence Berkeley National Laboratory.

56. Prindle, B., et al. (2007). "Quantifying the effects of market failures in the end-use of energy." Final Draft Report Prepared for International Energy Agency at http://www. aceee. org/energy/IEAmarketbarriers. pdf.
57. Rogers, E. M. (1962). Diffusion of innovations, Free Press of Glencoe.
58. Sanstad, A., et al. (2006). "End-Use Energy Efficiency in a 'Post-Carbon'California Economy: Policy Issues and Research Frontiers." Managing greenhouse gas emissions in California.
59. Savage, L. (1954). "The foundations of statistics Wiley." New York.
60. Schultz, W. P., et al. (2008). "Using normative social influence to promote conservation among hotel guests." Social influence 3(1): 4-23.
61. Shannon, C. E. (2001). "A mathematical theory of communication." ACM SIGMOBILE Mobile Computing and Communications Review 5(1): 3-55.
62. Shannon, C. E. and W. Weaver (1998). The mathematical theory of communication, University of Illinois press.
63. Shogren, J. F. and L. O. Taylor (2008). "On Behavioral-Environmental Economics." Review of Environmental Economics and Policy 2(1): 26-44.
64. Siew, L. W., et al. (2015). "The Impact of Human Behaviour Towards Portfolio Selection in Malaysia." Procedia - Social and Behavioral Sciences 172: 674-678.
65. Simon, H. A. (1955). "A Behavioral Model of Rational Choice." The Quarterly Journal of Economics 69(1): 99-118.
66. Sperling, D. and A. Eggert (2014). "California's climate and energy policy for transportation." Energy Strategy Reviews 5: 88-94.
67. Sperling, D. and A. Eggert (2014). "California's climate and energy policy for transportation." Energy Strategy Reviews 5: 88-94.
68. Su, Q. (2011). "The effect of population density, road network density, and congestion on household gasoline consumption in US urban areas." Energy Economics 33(3): 445-452.
69. Sutherland, R. J. (1991). "Market Barriers to Energy-Efficiency Investments." The Energy Journal Volume 12(Number 3): 15-34.
70. Train, K. (1985). "Discount rates in consumers' energy-related decisions: A review of the literature." Energy 10(12): 1243-1253.

71. Turrentine, T. S. and K. S. Kurani (2007). "Car buyers and fuel economy?" Energy Policy 35(2): 1213-1223.

72. Tversky, A. (1972). Elimination by Aspects: A Theory of Choice.

73. Varaiya, P. (2007). "Effectiveness of California's high occupancy vehicle (HOV) system." Unpublished, California PATH Research Report UCB-ITS-PRR-2007-5.

74. Wang, C. Y. and R. J. Carroll (1995). "On Robust Logistic Case-Control Studies with Response-Dependent Weights." Journal of Statistical Planning and Inference 43(3): 331-340.

75. Warren, R. A. and H. Simpson (1976). "The young driver paradox."

76. Zhou, J. (2012). "Sustainable transportation in the US: A review of proposals, policies, and programs since 2000." Frontiers of Architectural Research 1(2): 150-165.

77. December 2009 Dashboard: Year-End Tally". hybridCars.com. 2010-01-20. Archived from the original on 2010-01-22. Retrieved 2010-09-13.

78. February 2009 Hybrid Market Dashboard" (PDF). hybridCars.com. 2009-03-13. Retrieved 2010-03-15.

79. February 2008 Hybrid Market Dashboard" (PDF). hybridCars.com. 2008-03-14. Retrieved 2010-03-15.

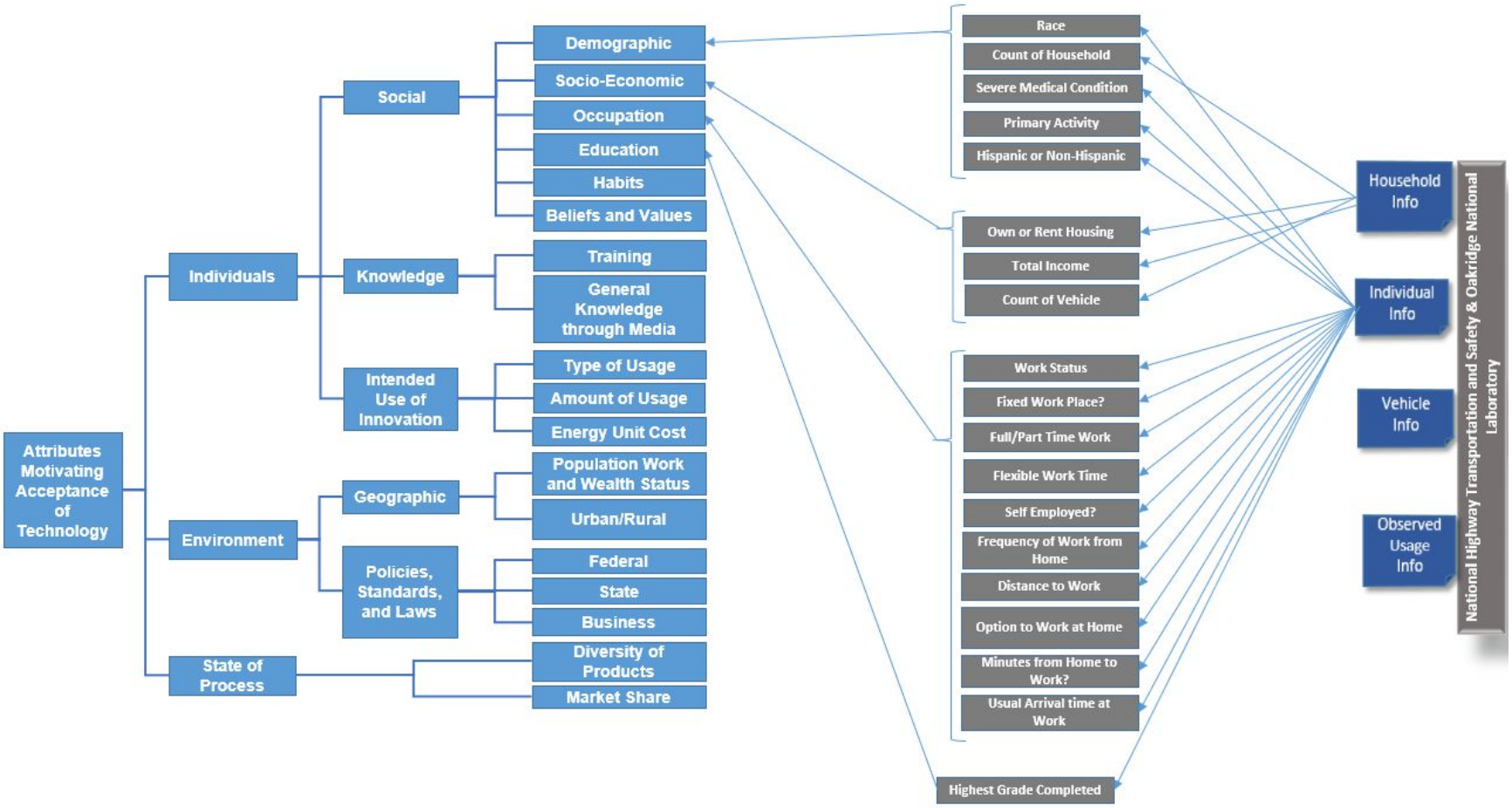
80. U.S. Department of Transportation, Federal Highway Administration, 2009 National Household Travel Survey. URL: http://nhts.ornl.gov.

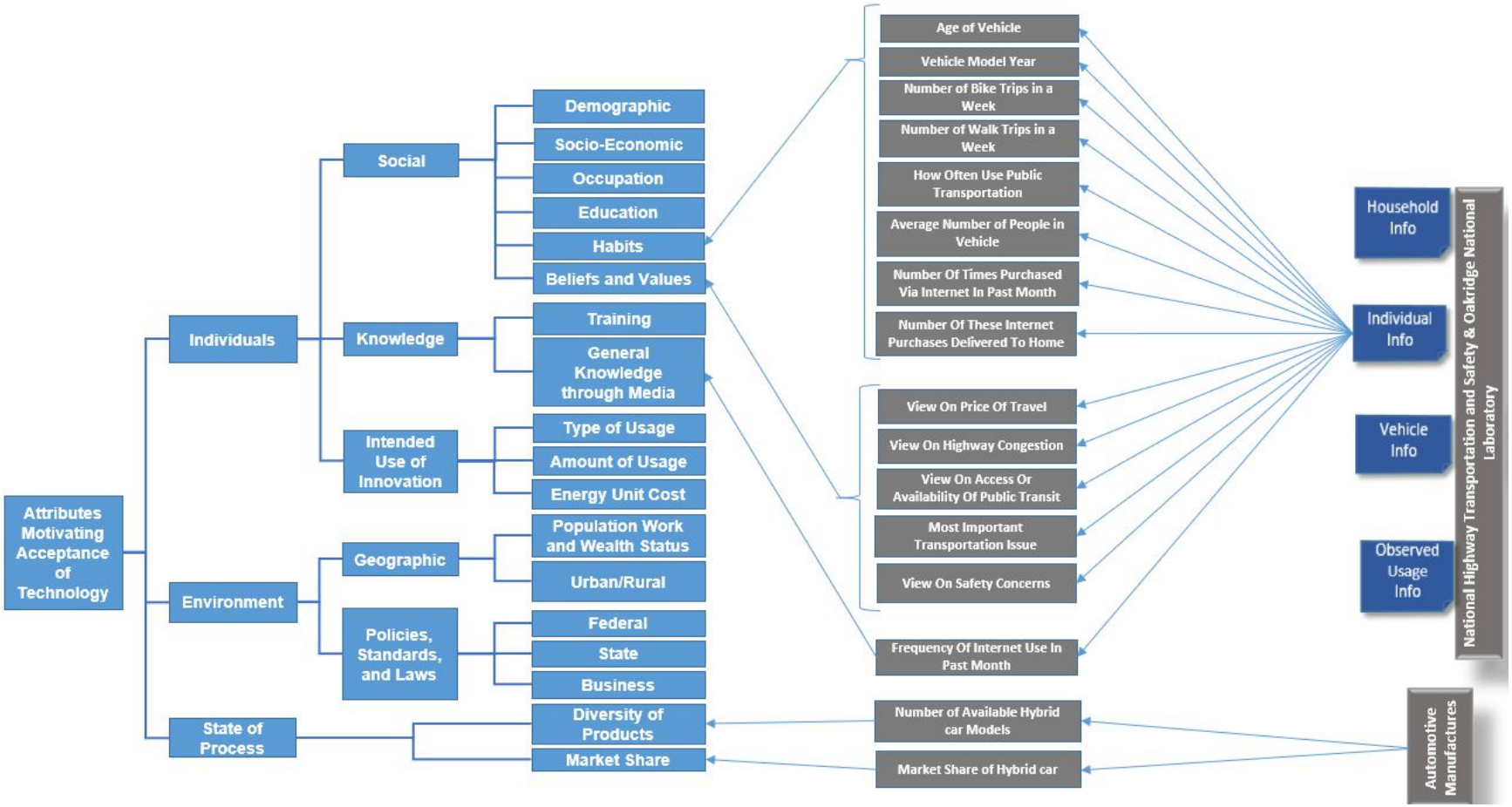
81. Shannon, C. (1948). A Mathematical Theory of Communication. Bell Labs Technical Journal, 27(4), 623-656.

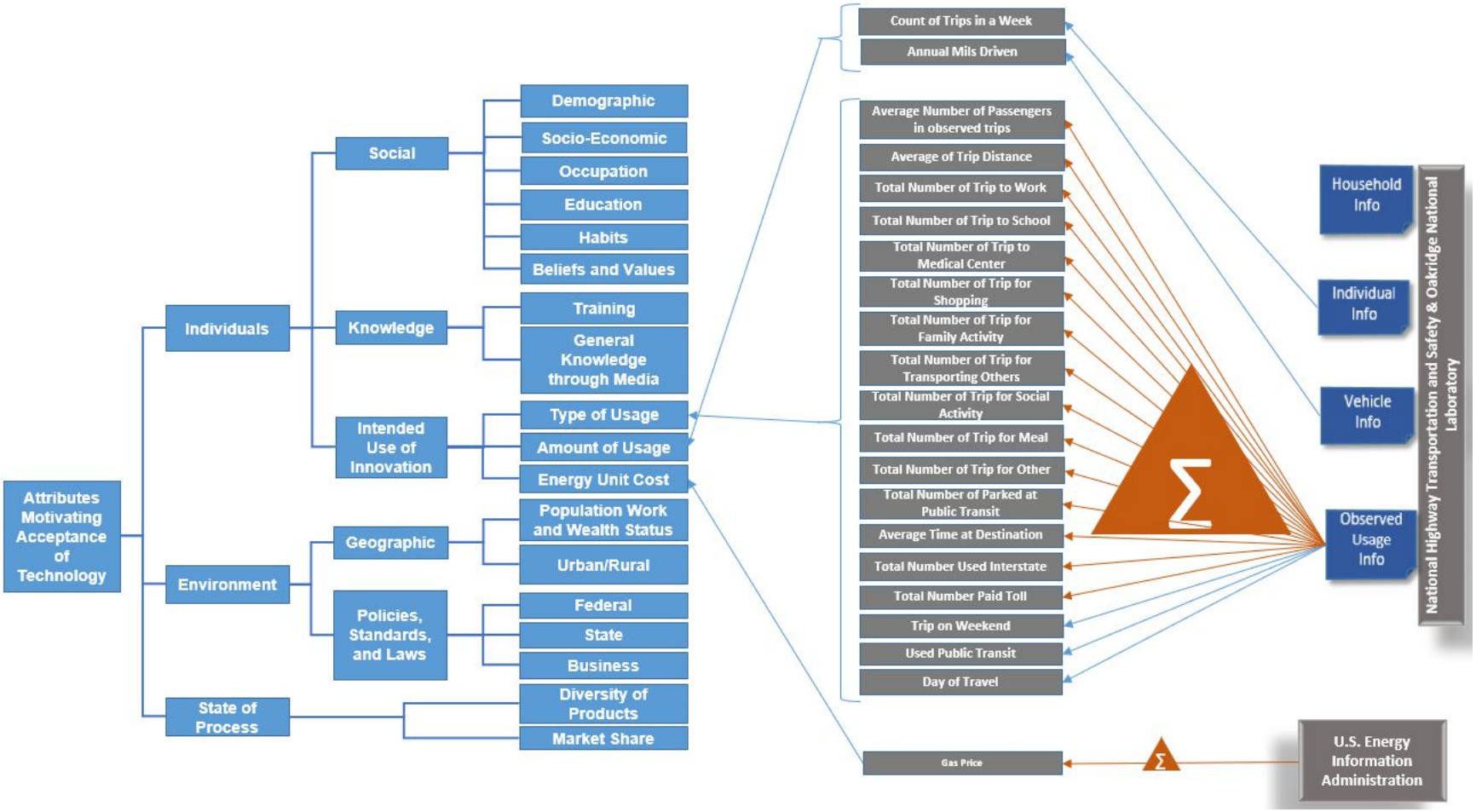
82. MLA style: "Daniel L. McFadden - Facts". Nobelprize.org. Nobel Media AB 2014. Web. 29 Jan 2018.

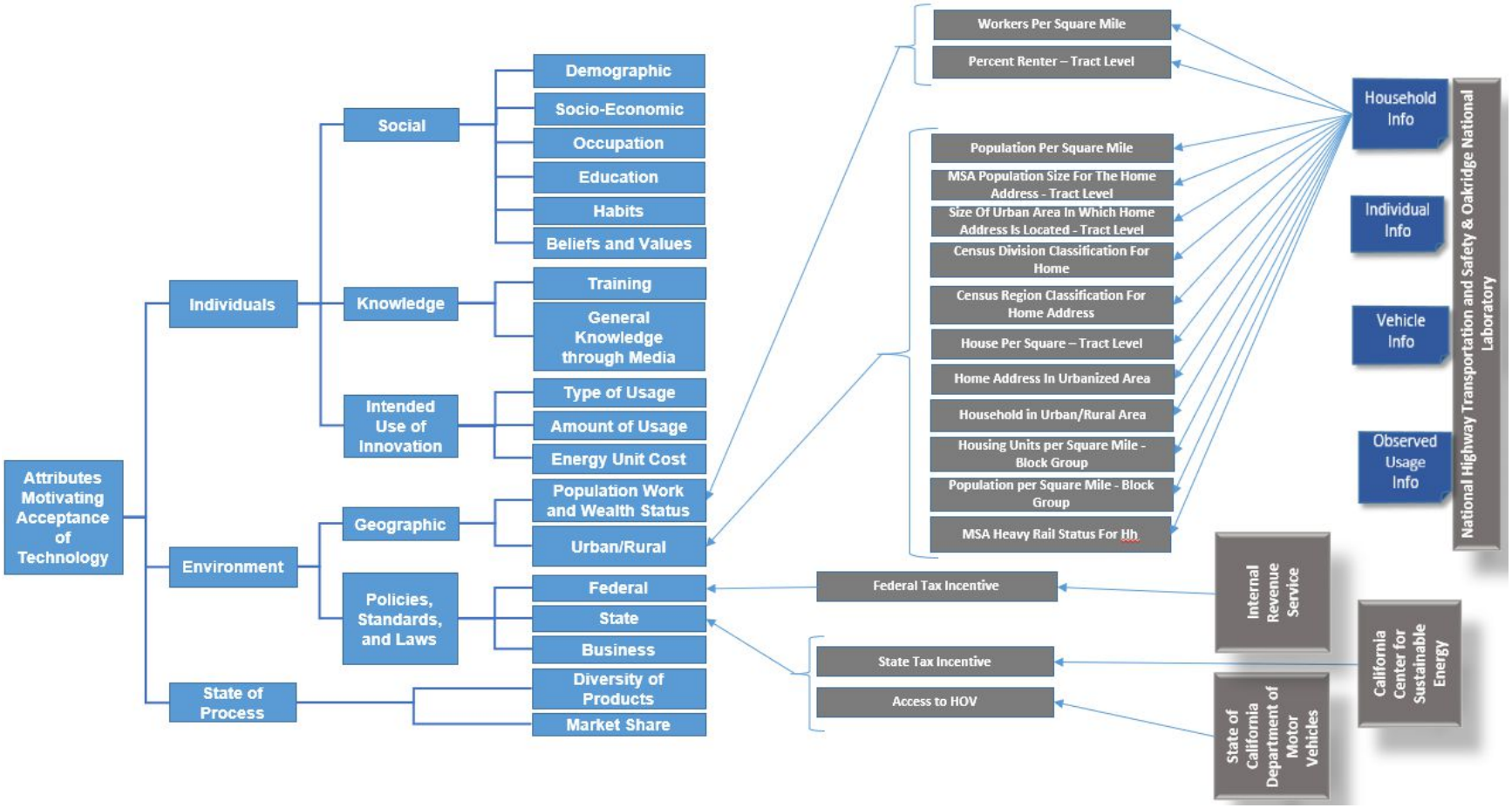
# Appendix

*Details of Extracted Attributes and their Source*




Attributes Motivating Acceptance of Technology
Individuals
Environment
State of Process
Social
Knowledge
Intended Use of Innovation
Geographic
Policies, Standards, and Laws
Demographic
Socio-Economic
Occupation
Education
Habits
Beliefs and Values
Training
General Knowledge through Media
Type of Usage
Amount of Usage
Energy Unit Cost
Population Work and Wealth Status
Urban/Rural
Federal
State
Business
Diversity of Products
Market Share
Age of Vehicle
Vehicle Model Year
Number of Bike Trips in a Week
Number of Walk Trips in a Week
How Often Use Public Transportation
Average Number of People in Vehicle
Number Of Times Purchased Via Internet In Past Month
Number Of These Internet Purchases Delivered To Home
View On Price Of Travel
View On Highway Congestion
View On Access Or Availability Of Public Transit
Most Important Transportation Issue
View On Safety Concerns
Frequency Of Internet Use In Past Month
Number of Available Hybrid car Models
Market Share of Hybrid car
Household Info
Individual Info
Vehicle Info
Observed Usage Info
National Highway Transportation and Safety & Oakridge National Laboratory
Automotive Manufactures


Attributes Motivating Acceptance of Technology
Individuals
Environment
State of Process
Social
Knowledge
Intended Use of Innovation
Geographic
Policies, Standards, and Laws
Demographic
Socio-Economic
Occupation
Education
Habits
Beliefs and Values
Training
General Knowledge through Media
Type of Usage
Amount of Usage
Energy Unit Cost
Population Work and Wealth Status
Urban/Rural
Federal
State
Business
Diversity of Products
Market Share
Count of Trips in a Week
Annual Mils Driven
Average Number of Passengers in observed trips
Average of Trip Distance
Total Number of Trip to Work
Total Number of Trip to School
Total Number of Trip to Medical Center
Total Number of Trip for Shopping
Total Number of Trip for Family Activity
Total Number of Trip for Transporting Others
Total Number of Trip for Social Activity
Total Number of Trip for Meal
Total Number of Trip for Other
Total Number of Parked at Public Transit
Average Time at Destination
Total Number Used Interstate
Total Number Paid Toll
Trip on Weekend
Used Public Transit
Day of Travel
Gas Price
Σ
Household Info
Individual Info
Vehicle Info
Observed Usage Info
National Highway Transportation and Safety & Oakridge National Laboratory
U.S. Energy Information Administration


Attributes Motivating Acceptance of Technology
Individuals
Environment
State of Process
Social
Knowledge
Intended Use of Innovation
Geographic
Policies, Standards, and Laws
Demographic
Socio-Economic
Occupation
Education
Habits
Beliefs and Values
Training
General Knowledge through Media
Type of Usage
Amount of Usage
Energy Unit Cost
Population Work and Wealth Status
Urban/Rural
Federal
State
Business
Diversity of Products
Market Share
Workers Per Square Mile
Percent Renter – Tract Level
Population Per Square Mile
MSA Population Size For The Home Address - Tract Level
Size Of Urban Area In Which Home Address Is Located - Tract Level
Census Division Classification For Home
Census Region Classification For Home Address
House Per Square – Tract Level
Home Address In Urbanized Area
Household in Urban/Rural Area
Housing Units per Square Mile - Block Group
Population per Square Mile - Block Group
MSA Heavy Rail Status For Hh
Federal Tax Incentive
State Tax Incentive
Access to HOV
Household Info
Individual Info
Vehicle Info
Observed Usage Info
National Highway Transportation and Safety & Oakridge National Laboratory
Internal Revenue Service
California Center for Sustainable Energy
State of California Department of Motor Vehicles

# Vita

Mohammad Ali Asudegi was born in Tehran, Iran, to parents Akbar Asudegi and Esmat Farivar. He attended Kamal High School and continued to Danesh High School in Tehran. He received his Bachelor of Science in Industrial Engineering in 2004. After five years of industry experience in Iran, he headed to University of Nevada, Reno, where he received a Master of Business Administration. He was working as Teaching Assistant in the Information Systems Department of the University of Nevada during his graduate study from 2010 to 2013. Ali received the award of Outstanding Graduate Student from the Department of Information Systems at the University of Nevada in 2013. Then, he accepted a Research Assistant position at the University of Tennessee, Knoxville, in the Industrial and Systems Engineering Department. Ali graduated with a Ph.D. in Industrial and Systems Engineering from the University of Tennessee, Knoxville, in 2018. He successfully finished many projects with local industries in the State of Tennessee in his role as a Research Assistant.