

# Analyzing and Predicting Heterogeneous Customer Preferences in China's Auto Market Using Choice Modeling and Network Analysis

Mingxian Wang and Wei Chen  
Northwestern Univ.

Yan Fu and Yong Yang  
Ford Motor Co.

## ABSTRACT

As the world's largest auto producer and consumer, China is both the most promising and complex market given the country's rapid economic growth, huge population, and many regional and segment preference differences. This research is aimed at developing data-driven demand models for customer preference analysis and prediction under a competitive market environment. Regional analysis is first used to understand the impact of geographical factors on customer preference. After a comprehensive data exploration, a customer-level mixed logit model is built to shed light on fast-growing vehicle segments in the Chinese auto market. By combining the data of vehicle purchase, consideration, and past choice, cross-shopping behaviors and brand influence are explicitly modeled in addition to the impact of customer demographics, usage behaviors, and attributes of vehicles. Scenario analyses are performed for segment demand forecasting by examining influencing factors such as economic change, fuel economy improvement and infrastructure development. Finally, a new network analysis approach is proposed to model customer cross-shopping behaviors that can inform the firm about the implied market structure and product competitive positioning. Our proposed approach is demonstrated by using a rich set of market data collected in China.

**CITATION:** Wang, M., Chen, W., Fu, Y., and Yang, Y., "Analyzing and Predicting Heterogeneous Customer Preferences in China's Auto Market Using Choice Modeling and Network Analysis," *SAE Int. J. Mater. Manf.* 8(3):2015, doi:10.4271/2015-01-0468.

## 1. INTRODUCTION

As individual incomes in China have risen dramatically, so have consumers' interests in cars that were previously out of their reach. In 2010, China overtook the United States to become the largest automotive market in the world. Market analysts estimate that vehicle sales will grow by a healthy 6 percent year-over-year through 2020 [1]. Auto manufacturers from around the world have China at the very center of their long-term strategies.

For foreign manufacturers to increase their market share in China, it is critical to understand the needs and preferences of Chinese consumers in different regions and segments. For example, how customers' expectations differ from region to region and from vehicle segment to segment. However, daunting challenges exist under the increasingly competitive and volatile market. First, auto manufacturers need to identify a diverse mix of values, preferences and behaviors that influence consumers' consideration and buying decisions. In response to these preferences, auto manufacturers then have to decide on the mix of products (vehicle models/segments) and technology content (e.g., fuel type, transmissions, engines, features, etc.) to be used for different regional markets.

In this paper, we focus on developing methods for modeling heterogeneous customer preferences and complex choice behaviors. The associated example of Chinese auto market is considerably more complex than other traditional applications, due to the existence of a large number of product offerings and complex customer profiles. Specifically, we aim to answer the following three questions that reflect three different aspects of customer preference:

*Q1) How do customer needs and preferences differ from region to region? What are possible main market forces that shape these regional differences?*

*Q2) How to model and predict customer preferences under various impact factors, e.g. rising income, changing price, increased fuel economy, change in driving behaviors, etc.?*

*Q3) How to derive the market segmentation and vehicle competition based on customers' consideration decisions?*

Particularly in mature markets, extensive data sets are available for predictive analytics on customer segmentation, demand prediction, cost and profitability analysis, and product optimization. Many advanced technologies for data collection have emerged including web page tracking, in-car sensors, GPS tracking, and more, to help

facilitate the process of data collection. The main power of data analytics is to enable auto manufacturers to derive value from the extracted data. Based on this data, predictions about customer preference behaviors can be generated for making rigorous decisions that were once decided with gut instinct. For example, there is growing interest in *sentiment analysis* [2, 3], which utilizes information from publicly available data sources like online reviews and forum discussions. Using text analytics coupled with powerful machine learning, high-quality results can be generated to uncover early-stage preference trends and crowd sourced customer feedback for product improvement. As another example, a company called *Autometrics* predicts vehicle demand by tracking the web browsing habits of potential car buyers from 150 different third-party car-buying websites. The up-to-the-minute information is then sold to major auto manufacturers to help adjust production schedules, tweak marketing campaigns and minimize inventory [4]. Although the automotive industry is more data-driven today, building a preference demand model is still a difficult task as it inherently relates to modeling human behaviors.

In this paper, aggregated statistical analyses are first conducted to develop insights into regional preference differences (*Q1*). Early work on customer preference modeling can be traced back to marketing and econometrics literature where various methods are developed, such as Multiple Discriminant Analysis, Factor Analysis, Multidimensional Scaling, Conjoint Analysis, etc [5]. Related to geographical variables, choropleth maps are normally used to display a regional variable on the map by producing virtually continuous shades in proportional to the measurement of the examined variable [6]. Using this technique, one can visualize how a group measurement varies across a geographic area and the level of variability within a region. However, the aforementioned techniques are aggregated methods that only focus on group averages of macroeconomic variables. As a result, they are insufficient to model preference heterogeneity.

With the second goal of forecasting heterogeneous customer preferences under various impact factors (*Q2*), disaggregated preference models are needed to drill down into individual customer choice behaviors instead of group averages for demand prediction. As opposed to the aggregate studies, Discrete Choice Analysis (DCA) [7, 8] uses data of individual customers instead of group averages to explain why an individual makes a particular choice given his/her circumstances, and therefore it predicts changes in choice behavior due to changes in individual characteristics and product attributes. Using DCA to estimate demand entails estimating choice probability for a given product alternative, and aggregating choice probability for a given product to estimate its choice share, and ultimately its demand.

Following the seminal work of Ben-Akiva and Lerman [9, 10] and Lave and Train [11], a wide variety of discrete choice models have been developed in the automotive demand literature. Representative work includes the hedonic (random coefficient) logit model proposed by Boyd and Mellman [12] in which variations in taste coefficients of vehicle attributes are first incorporated, but the customer attributes are still missing. The BLP approach proposed by Berry, Levinsohn and Pakes [13] enables one to handle the endogeneity through standard instrumental variables estimation, but the method requires a

very large number of constants in the choice model which may cause numerical issues in estimation. Using mixed logit formulation, Train et al. proposed several vehicle purchase choice models to predict customers' preferences for alternative fuel vehicles [14, 15, 16, 17]. Their work showed that mixed logit provides a flexible and computationally practical approach to discrete choice analysis. Other applications of discrete choice models for vehicle demand are essentially variations of the preceding literature. Also, the engineering community has implemented and extended DCA broadly to model demand for an automotive system, in particular in the context of product design optimization [18, 19, 20, 21, 22, 23, 24, 25]. Built on the extant literature on preference demand modeling, in this paper, we use mixed logit to examine the underlying drivers of consumer heterogeneity in vehicle preferences.

However, there are several challenges for applying the utility-based DCA in practical applications. One fundamental obstacle is its oversimplified assumption based on Rational Choice Theory [26]. That is, DCA only captures how consumers make trade-offs on product attributes among a competing set of products, but ignores the structure of the choice set and the complex relationships among the competing products that imply the consideration decisions of customers. As the "choice set" is not treated as a part of the compensatory decision making process the consumer uses to arrive at choice on a specific occasion, it is not explicitly included in a traditional utility formulation [27].

To handle the limitation of DCA in terms of the customer consideration decisions, we propose a novel network modeling approach as a compensatory approach separated from DCA to identify the association relationships between products based on the customer consideration and purchase data (*Q3*). The key idea of this approach is to construct a network graph that contains hundreds of product alternatives (vehicles) in the market as network nodes. The links between two products (nodes) reflect the proximity or similarity of two products in the customers' perceptual space, indicating what product alternatives are more likely to be considered together by a customer. In literature, similar strategy has been adopted to generate customer perceptual maps in marketing research [28] and improving the mining strategy of the transactional data in information system research [29, 30]. In this work, the method of association networks is extended to represent the relational pattern of products in consideration and choice data, and a graphical approach is used to visualize and analyze the network structures for future product planning.

The reminder of the paper is organized as follows. [Section 2](#) introduces the dataset used for the complete study. [Section 3](#) describes the regional statistical analysis on aggregated customers' behaviors but using granular lenses. [Section 4](#) presents the mixed logit discrete choice model for preference analysis and scenario prediction. [Section 5](#) proposes two network models that were built to facilitate the segmentation and ranking of products based on customer consideration and choice decisions. [Section 6](#) concludes the paper by comparing the different methods and raising discussion topics for future research.

## 2. THE DATA SET

In this paper, we use the New Car Buyers Survey (NCBS) data collected by an independent research company (IPSOS) in 2013 to study the customer preferences in China's auto market. The survey is initially conducted to obtain opinions from new car drivers on a number of issues, such as the subjective quality of the new vehicle, problems occurred during usage, vehicle delivery experience etc. To ensure the respondents have adequate user driving experience, all respondents are required to own the car no less than 6 weeks and no more than 32 weeks. The data set contains rich information of a diverse set of factors, including customers' new vehicle choice by make and model, ranking of the vehicles they seriously considered acquiring, previous owned vehicle and other vehicles in household, as well as customer socio-demographic and usage characteristics. Vehicle attributes, such as purchase price and fuel consumptions, are also reported by customers in the survey. For the purpose of preference modeling, we only focus on those sampled respondents who identified themselves as the major decision maker in vehicle purchase as well as the main driver in subsequent vehicle use. This prescreening results in a subset data of 49921 customers and 389 vehicle models for analysis and prediction.

## 3. REGIONAL STATISTICAL ANALYSIS

Similar to the trend of urbanization, China's car market has strong regional characteristics. Factual data shows that, from 2002 to 2011, China's smaller cities (Tier 3/4 cities) contributed about 40 percent of total new car sales. However, city tiers alone are not sufficient to describe geographical differences in new car choices. To address research question one (Q1), "How do customer needs and preferences differ from region to region?", we begin by investigating the geographical information of customers to determine the degree of preference variations across different regions.

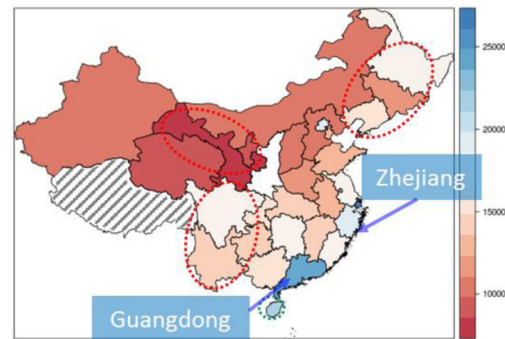
Although China has experienced remarkable economic growth as a whole, significant regional variation exists. To see the impact of economic drivers on preference disparities, we look for factors that may be indicators of local economic status. In Fig. 1, customers' income is examined against the purchased vehicle price. Our estimates are based on the province-level statistics such that the number of buyers from each province is visualized using bi-polar color progressions in the geographical map (also known as choropleth map). The shades of color are based on the absolute value of the quantity of interest, where blue refers to high range and red represents low range.

Disparity has been observed across different regions. As seen in Fig. 1, those living in the coastal provinces of Zhejiang and Guangdong are more willing to spend money on premium vehicles. This might be because this region has had more exposure to the premium car market for several years. Customers might view owning a premium car as an indicator of social status and lifestyle. It is also noted that consumers in Southwest China, Northeast China, and West Central China generally are more willing to spend money on vehicles, even though the incomes in these regions are not the highest in the country. Those consumers are more likely to become potential buyers for luxury higher-end vehicles when they have sufficient incomes. In contrast,

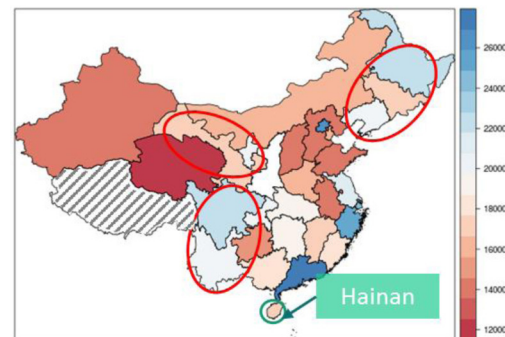
consumers from Hainai province appear to be mostly budget buyers who are more price-sensitive, even though the income in this region is high.

It is also found that due to the differences in infrastructure development and geographical conditions, driver's vehicle usage behaviors differ dramatically across regions. For example, off-road driving is still predominant in some less-developed provinces in Central China (e.g., Ningxia and Gansu) and South Central China (e.g., Hunan and Guizhou), where vehicle overall usage in terms of driving distance is low, possibly due to the severe road conditions.

Apart from the differences in economic status and infrastructure development, local cultures and social influence might have shaped customer vehicle preferences as well. It is interesting to note that there is no strong correlation between SUV purchase and offroad driving needs. Customers choose SUV for other reasons such as trendiness, tastes or personality, safety, roominess, taller seating, better visibility, pulling capacity, passenger seating, cargo space, etc.



a. Monthly Household Income (in RMB)



b. Vehicle Purchase Price (in RMB)

Figure 1. Comparison between Household Income and Vehicle Price. Data: NCBS 2013.

Through regional statistical analysis, we have gained an initial understanding about the impact of customer social-economic status and usage behaviors on vehicle preference choices. Beyond this, we are interested in quantitatively assessing these impact factors for demand prediction. To achieve the goal, discrete choice analysis is employed in the next section.