2015-01-0468
Published 04/14/2015

doi:10.4271/2015-01-0468
saematman.saejournals.org

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

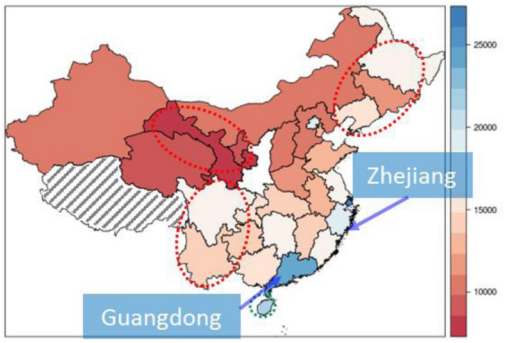
Similar to the trend of urbanization, China's car market has strong regional characteristics. Factual data shows that, from 2002 to 2011, China's smaller cities (Tier 3/4 cities) contributed about 40 percent of total new car sales. However, city tiers alone are not sufficient to describe geographical 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.

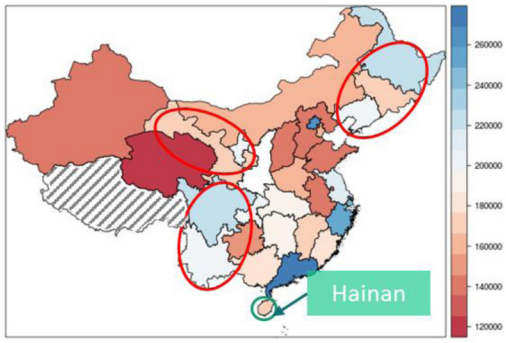
Disparity has been observed across different regions. As seen in Fig. 1, those living in the coastal provinces of Zhejiang and Guangdong are more willing to spend money on premium vehicles. This might be because this region has had more exposure to the premium car market for several years. Customers 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.

4. DISCRETE CHOICE ANALYSIS

As discussed in Section 1, Discrete Choice Analysis (DCA), being a disaggregated approach, can generate more useful and accurate predictions when evaluating the impact of changing market, different target consumers, and various product configurations. In this section, the technical background of discrete choice models is first introduced, followed by a mixed logit model example for China market.

4.1. Technical Backgrounds of Multinomial Logit and Mixed Logit

As a probabilistic approach, DCA formulates an intermediate utility function to model choices customers make among competing products and predicts the probability that a product is chosen over others. Following the utility maximization theory, the individual choice utility U for product alternative i and individual customer n consists of a deterministic part W and a random part ε, given a set of competing products available to the customer (named as "choice set") CS_n.

$$U_{in} = W_{in} + \varepsilon_{in} \quad \text{if} \quad i \in CS_n \,, \tag{1}$$

The observed utility W is further parameterized as a function of customer-desired attributes **A**, socio-demographic attributes **S**, and unknown coefficients β which can be estimated given the actual choices customers make.

$$W_{in} = f\left(\beta : \mathbf{A}_{in}, \mathbf{S}_{in}\right), \tag{2}$$

In the multinomial logit, the preference coefficient (β) for the product attributes (**A**) is assumed to be identical across all customers. That means, customers' differences in choice preferences is only expressed by considering customer socio-demographical attributes (**S**), such as gender, age, income, etc. Assuming the utility function is a linear combination of attributes, the observed utility then has the following structure:

$$W_{in}\left(MNL\right) = \boldsymbol{\beta} \cdot \mathbf{Z} = \beta_i + \boldsymbol{\beta}_S \mathbf{S}_n + \boldsymbol{\beta}_A \mathbf{A}_i + \boldsymbol{\beta}_{AS}\left(\mathbf{A}_i \mathbf{S}_n\right), \tag{3}$$

Z is the set of all attributes containing **A** and **S**. The β_i is a constant specific to each product alternative i (named as Alternative Specific Constant), which captures the average effect on utility of all factors that are not included in the model. Conversely, the Alternative Specific Variables ($\boldsymbol{\beta}_S$) are coefficients for socio-demographic variables relating to the customer that do not vary over the alternatives. They are utilized to represent the heterogeneity of customer preferences for each alternative due to the differing socio-demographic attributes **S**. The coefficients $\boldsymbol{\beta}_A$ evaluate the importance of various vehicle attributes and the coefficients $\boldsymbol{\beta}_{AS}$ capture the interactions between customer attributes and vehicle attributes.

The choice probability in MNL is derived by assuming the error terms ε_{in} are Independently Identically Distributed (IID), following a Gumbel (Type I extreme value) distribution.

$$\Pr_n\left(i : CS_n\right) = \frac{e^{W_{in}(\beta)}}{\sum_{j=1}^{J} e^{W_{jn}(\beta)}} \,, \tag{4}$$

In the above equation, $\Pr_n(i{:}CS_n)$ refers to the probability of choosing alternative i from choice set CS_n available to customer n; j represents any alternative in choice set such that $j \in CS_n$.

As the simplest form of choice models, MNL has at least two major limitations [7]. First, MNL is not immune from the property of Independence of Irrelevant Alternatives (IIA), which leads to the proportional substitution patterns among the alternatives considered. The IIA property is often not a realistic assumption and is not desirable for most real applications. Second, MNL models imply that the preference coefficient (β) is fixed across customers. Thus, the model can only capture taste variation for observed **S** attributes (systematic heterogeneity), but not differences in taste that cannot be linked to **S** (random heterogeneity).

Distinguished from MNL, the issues of IIA and random heterogeneity can be appropriately handled if a MXL formulation is applied. The observed utility follows

$$W_{in}\left(MXL\right) = \boldsymbol{\beta} \cdot \mathbf{Z} = \boldsymbol{\beta}_{A,n} \mathbf{A}_i + \boldsymbol{\beta}_S \mathbf{S}_n + \boldsymbol{\beta}_{AS}\left(\mathbf{A}_i \mathbf{S}_n\right), \tag{5}$$

In MXL, each individual person has his/her own set of model coefficients $\boldsymbol{\beta}_{A,n}$ for product attributes $\mathbf{A}_i$. Coefficients $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_{AS}$ are normally fixed over customers, but can be set to random distributions depending on the level of analysis. The choice probability is then expressed as integrals of the MNL probabilities over distributions of random coefficients.

$$\Pr_n\left(i : CS_n\right) = \int \left(\frac{e^{W_{in}(\beta)}}{\sum_{j=1}^{J} e^{W_{jn}(\beta)}} \right) f\left(\beta\right) d\beta \,, \tag{6}$$

where $f(\beta)$ is the probability density function for all random components in $\boldsymbol{\beta}$. It can be shown that this specification relaxes the IIA restriction that allows the existence of correlations between alternatives.

Estimation of discrete choice models can be accomplished using the Maximum Likelihood approach, in which the parameters β are determined through estimation of the log-likelihood function LL

$$LL = \log L = \sum_{n=1}^{N} \sum_{i=1}^{J} y_{ni} \log\left(\Pr_n\left(i\right)\right), \tag{7}$$

where y_{ni} is the binary choice indicator that equals 1 if customer n chooses alternative i and 0 otherwise. Although MXL is far more general and flexible than MNL, there is no closed form solution for Eqn. (7). In practice, numerical simulation is often conducted by taking a finite number of draws from the distribution of random coefficients to approximate the mixed logit choice probability.

4.3. Mixed Logit Preference Model for China Market

Construction of the mixed logit model is complicated by the needs of incorporating various types of information in the dataset including customer's new vehicle choices by make and model, customer socio-demographical and usage characteristics, vehicles that customers seriously considered acquiring, customer ratings towards various vehicle aspects, and the vehicle ownership history. Before specifying the model structure, we first describe how we take into the account the following three aspects in our modeling: 1) customer consideration decisions (or cross-shopping decisions), 2) customer subjective ratings, and 3) the effect of brand influence and brand loyalty on vehicle choices.

4.3.1. Issues in Mixed Logit Construction

1) Customer Considerations Decisions

Understanding customer's consideration decisions is critical to modeling customer preference heterogeneity. It has been shown in literature that an accurate estimation of customer consideration decisions will often improve the accuracy of choice prediction [31]. Although NCBS does not include choice set information specifically, respondents are asked to list sequentially other vehicles that they seriously considered, from the most relevant to the least relevant, in addition to the vehicle they purchased.

To integrate customer consideration decisions into choice model, we follow the preference ranking approach proposed by Train and Winston [16]; the approach was applied to research the declining of 2009 U.S. auto market. The principle of the approach is to capture customers' consideration and choice decisions together by the value of utility. It is assumed that a considered vehicle should have a utility value lower than the purchased vehicle, but higher than other vehicles that are irrelevant to the decision problem. As an example, for customers who indicated only one considered vehicle *h*, and purchased vehicle *i*, we have:

$$U_{in} > U_{hn} > U_{jn}, \qquad \text{for all} \quad j \neq i, h, \tag{8}$$

When the unobserved error component follows I.I.D. extreme value distribution, the probability of a utility ranking is a product of logit formulas [32]. Therefore, we can derive the probability that a consumer buys vehicle *i* and also considered vehicle *h* as

$$\Pr_n(i,h) = \int \left(\frac{e^{W_{in}(\beta)}}{\sum_{j=1}^{J} e^{W_{jn}(\beta)}} \right) \left(\frac{e^{W_{hn}(\beta)}}{\sum_{j=1, j\neq i}^{J} e^{W_{jn}(\beta)}} \right) f(\beta) d\beta, \tag{9}$$

For customers who indicated more than one considered vehicles, we can generate a utility ranking in the order that they mentioned the vehicles in survey. As shown in literature, one advantage of the utility ranking approach for mixed logit models is that only when the consideration decisions are explicitly included in the model, the standard deviations for several random β coefficients are found to be significant [16, 33]. In contrast, when consideration decisions are ignored, the unobserved taste variations cannot be correctly estimated.

2) Customer Subjective Ratings

Besides the traditional vehicle attributes, vehicle purchase behavior is also highly influenced by qualitative considerations, such as customer's perceptions of the exterior styling and their impression of overall quality of the vehicle. In NCBS data, the qualitative considerations are recoded using discrete rating scores by asking customers related questions, considering overall quality and design as well as distinct aspects of different vehicle subsystems, components and features. A 1-10 point satisfaction rating is used to represent respondents' opinions from *completely dissatisfied*, *somewhat dissatisfied*, *fairly satisfied*, *very satisfied*, to *completely satisfied*. As the collected rating information explains a large portion of preference heterogeneities, there is a need to incorporating such data in the discrete choice model.

Nevertheless, there are three major obstacles in using the rating data for choice modeling. First, as rating is a subjective measure reported by survey respondents, the data inevitably possesses measurement errors, bias, and scale usage disparities [34, 35] from heterogeneous respondents. Second, as customers were asked to rate in more than 40 questions and many of the questions are related, including all survey ratings directly in choice model would introduce undesirable correlations. High correlations between explanatory attributes may result in redundancy and suppression of model estimators, and run into the problem of model convergence. Another challenge encountered by most survey data is the lack of ratings for choice set alternatives. In NCBS, customer ratings are only given for the purchased vehicle. However, due to the unique structure of choice model, ratings for other alternatives in the choice set are needed for utility calculation.

To incorporate rating data in modeling and at the same time alleviate the above difficulties, we refer to the approach in [24]. To deal with the high correlation of rating responses, the 40+ ratings are grouped into five categories - *Styling*, *Safety*, *Quality*, *Handling* & *Comfort*. For each customer, the averaged rating score is then calculated as the categorical rating. The issue of missing ratings is handled by estimating categorical ratings for other choice set vehicles as a function of customer attributes, $\mathbf{S}$, with parameters, $\boldsymbol{\lambda}_{0,i}$ and $\boldsymbol{\lambda}_S$:

$$\hat{\mathbf{R}}_{in} = g(\mathbf{R},\mathbf{S}) = \boldsymbol{\lambda}_{0,i} + \boldsymbol{\lambda}_S \mathbf{S} \;\text{ where }\; \boldsymbol{\lambda}_{0,i} = \bar{\mathbf{R}}_i, \tag{10}$$

As seen in the model structure, the rating bias can be moderately corrected by the average satisfaction $\boldsymbol{\lambda}_{0,i}$, while the individual heterogeneity still persists as introduced by the demographical coefficients $\boldsymbol{\lambda}_S$. The actual and the predicted categorical ratings are used together as inputs for the mixed logit utility model.

3) Brand Effect

The issue of brand influence has consistently drawn interest from researchers in marketing community [36, 37]. As brand value cannot be directly attributed to the physical component of product, measuring the value created by such implicit factors as brand name associations are critically important in specifying discrete choice models.

We construct two indicators of *brand performance* to capture brand influence on different types of buyers - first-time buyers and loyal buyers. We use the brand information of purchased vehicles to create dummies. For first-time buyers, the brand dummies are brand indicators of one's current purchase. These variables are generated to capture the brand awareness, advertising and other unobserved heterogeneity like the sentiment around the brand message and how the target audience perceives the brand. For repeat buyers, the associated brand dummy equals 1 if the customer selects the same brand as his/her previous purchase, or the chosen vehicle brand is the same as any of other vehicles in household. We attempt to use the repeat buyer brand dummies to capture a customer's confidence and loyalty for that manufacturer. When a customer's ownership builds positive experience with a manufacturer, it is more likely that the customer will purchase that brand again. Nevertheless, the repeat buyer brand dummies could also capture other unobserved factors, like the variety of product offerings of a brand and customer's accessibility to other brands.

To better incorporate the heterogeneity of brand culture and marketing strategies, we also specify *brand origins* for American, European, Japanese, Korean, and Chinese manufacturers as additional attributes in the mixed logit model. In summary, the brand effect is captured by the coefficients (β) of both the *brand performance* dummies and *brand origins* in our choice model.

4.3.2. Model Specification

The mixed logit model is developed to integrate the customer's vehicle choices, the rankings of vehicles they considered, brand effects generated from vehicle choice histories, vehicle attributes, vehicle attributes ratings, customer socio-demographical and usage attributes. Hence, the deterministic utility of the mixed logit model is expressed as:

$$W_{in} = \boldsymbol{\beta}_{A,n}\mathbf{A}_{in} + \boldsymbol{\beta}_{R,n}\mathbf{R}_{in} + \boldsymbol{\beta}_{B}\mathbf{B}_{in} + \boldsymbol{\beta}_{AS}\left(\mathbf{A}_{in}\mathbf{S}_{n}\right), \quad (11)$$

Compared to the basic DCA structure in Eqn. (5), the above specification also accounts for customer subjective ratings $\mathbf{R}_{in}$ and brand effects $\mathbf{B}_{in}$ as we have previously described in Section 4.3.1. The coefficients of the (predicted) rating attributes $\boldsymbol{\beta}_{R,n}$, similar to that for the product attributes, can also be treated as random distributions over customers. In this study, we assume all random coefficients are independently normally distributed. In Eqn. (11), the alternative specific constants are omitted, because the vehicle (alternative) offerings are less likely to remain unchanged over the time. Even though the ignorance of alternative specific constants may deteriorate the model fitting, it allows for prediction of new vehicles demand in the future market [38].

In addition to the customer attributes, ratings, and brand effect attributes described earlier, the vehicle attributes $\mathbf{A}_i$ considered in this study include vehicle's purchase price, fuel consumption, horsepower, engine size, fuel type, drivetrain type, transmission type, body-type segments, and model origins (whether the car is imported or produced locally). Regional attributes are not included, because our model estimation tests show that all regional attributes we examined are not significant under the MXL specification. One explanation could be that the regional heterogeneity has been captured by the differences of other socio-demographical and usage attributes (e.g., income, type of road driving), which are included in the model. As the regional attributes may provide additional information for auto manufacturers, future work is needed to address the challenges of utilizing regional attributes in model prediction for new market.

As the choice probabilities (Pr_n) in Eqn. (9) are integrals with no closed forms, we use the simulated log-likelihood maximization approach to obtain the model parameters $\boldsymbol{\beta}_n$. We find that 70 out of 75 included attributes in the model are significant at 0.05 level, and most of them have expected signs, e.g., negative for price/income and fuel consumption, positive for turbocharged-engine dummy and automatic transmission dummy. Although the mean values of comfort and quality ratings are negative, both of them have significantly large standard deviations. One explanation is that there exists a large variation among respondents' preferences and some of the respondents care less on a vehicle's comfort and quality. Interestingly, the coefficient for styling is considerably higher than other rating coefficients. That means that, to most Chinese customers, a vehicle's styling is generally more important than other aspects when shop for new vehicles.

Table 1. Selected Coefficients in the Mixed Logit Choice Model. Significant Coefficients are shown in Bolds.

Random Elements Rel. to Vehicle Attributes	Mean	STD
Diesel dummy	**-2.632**	**1.456**
Alternative fuel (1 if the vehicle is powered by alternative-gas, electric or hybrid fuels, 0 otherwise)	**-2.937**	**1.157**
Turbocharger dummy	**0.824**	0.021
Automatic transmission dummy	**0.620**	0.082
Fuel consumption (in L/KM)	**-0.135**	**1.708**
Random Elements Rel. to Customer Ratings	**Mean**	**STD**
Safety	**0.188**	**2.144**
Comfort	**-0.210**	**1.472**
Styling	**0.852**	**1.897**
Quality	**-0.241**	**0.992**
Random Elements Rel. to Vehicle & Customer Attr.	**Mean**	**STD**
Price/Income (vehicle purchase price divided by monthly household income)	**-0.088**	**0.043**

In this study, we do not explicitly include macroscopic factors such as fuel prices, unemployment rate, and others to estimate future vehicle demands. The reason is that we are looking at choice data in a short time period (single year) and most of the macroscopic variables do not change dramatically. Nevertheless, our model is not completely insulated from macroeconomic issues. For example, if the economy declines and unemployment rate increases, customers may be more sensitive to car prices and thus the model's price/income coefficient will decrease. If the fuel price rises, customers would be more opt to purchase fuel-efficient cars, and the fuel consumption coefficient in the model will decline further accordingly.

4.3.3. Scenario Analysis

With the developed MXL model, scenario analysis is conducted to guide an exploratory evaluation of the ongoing increasing demand in China's auto market. In this paper, the scenario of income growth is presented using the sampled respondents in survey as an example to demonstrate

the capability of scenario analysis, whereas other impacting factors (e.g., SUV price, fuel efficiency, etc.) on different customer populations can be examined as well for target market prediction.

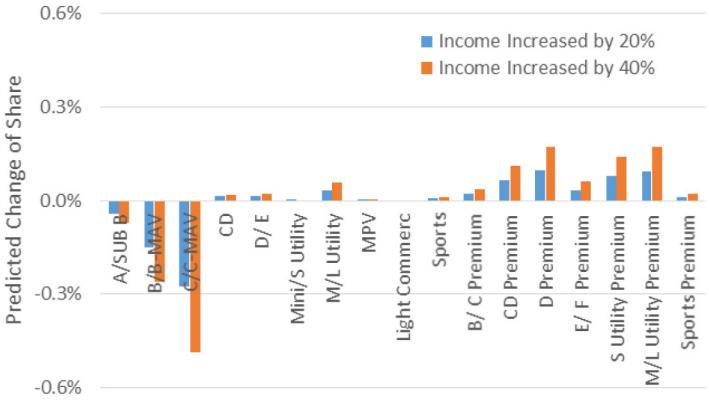


Figure 2. Scenario Analysis of Increased Income

To study the effect of increased income, we assume that every customer's income is increased by 20% and 40%, respectively. The choice probabilities are updated based on the new customer profiles, and then aggregated to derive the predicted demand for each vehicle segment. In Fig. 2, we plot the vehicle segments and the predicted change of market share under the examined scenarios.

When the customer income increases, the shares of the compact vehicles (segments below C) decreases. C/CMAV segment (e.g. VW Golf, Ford Focus, Chevrolet Cruze, Honda Civic) losses the most share, but this will not affect its largest share position in market. All premium segments are strongly affected and all their market shares increase significantly. By comparison, there is very little effect to mini/S SUV, MPV, and light commercial vehicle buyers.

While discrete choice models have many advantages, they lack a systematic way to model vehicle competition, and the modeling results may not fully explain the preference heterogeneity in vehicle consideration. In the next section, a new network approach is introduced to address the above needs. Different from discrete choice models that focus on choice decisions, we utilize the set of vehicles that customer's considered in cross-shopping activities to derive product competition maps and analyze more closely the consideration preference heterogeneity.

5. NETWORK MODELS

Network analysis has emerged as a key method for statistical analysis of complex systems in a wide variety of scientific, social, and engineering domains. The main benefit of this approach is its capability of visualizing complex relations in a simple network graph, where nodes represent individual members and ties represent relationships between members. Unlike traditional social network analysis that views social relations in terms of network theory, in this work we study the association relations between non-human technological artifacts (vehicles) and use the topological information of network to explain human's preference behavior.

Our research objective is to explore descriptive network analysis for use in identifying aggregated product associations and hierarchical relations. We propose to study two types of product association networks: a *Vehicle Association Network* formed by unidirectional links using customer consideration data, and a *Hierarchical Preference Network* with directed links generated by consideration data and choice data. The following sections demonstrate how to use the structural results from the two networks to determine the implied market structure, product positioning, product competition, and opportunities for new product launching.

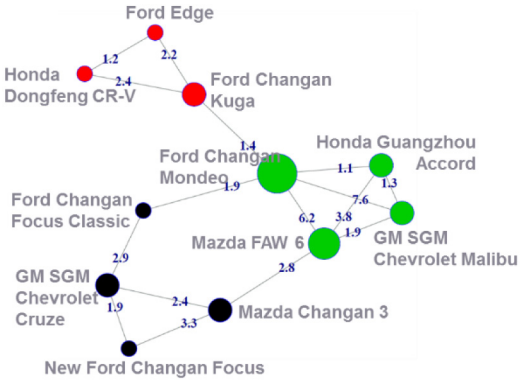
5.1. Technical Background of Association Networks

In a marketing study, Netzer et al. [28] employed a network analysis framework together with text mining to understand customers' top-of-mind associative network of products based on large-scale, customer-generated data posted on the Web. Although similar approaches are employed in our study, the new contribution of this work is on constructing vehicle networks based on both consideration and choice relations in marketing surveys for understanding customer preferences and product competition.

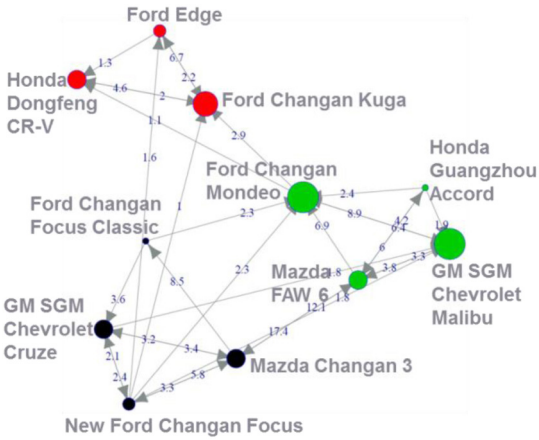
In creating the association network, links between two vehicles reflect the proximity of the two vehicles in a customer's mind when shopping for a new vehicle. The strengths of links are quantified using standard measures of association rules, such as the Lift [39], Jaccard index [40] and Cosine similarity [41], to reflect how often they are compared by a population of customers.

With the constructed association network, structural metrics can be assessed, even though their implications are different. **Centrality** [42, 43] measures a product's competitiveness, indicated by its level of connectivity to other products. We assume that more central (or more connected) products have higher levels of survivability in market competitions as a result of its structural advantageous. For the illustrative network in Fig. 3(a), Mondeo is more "central" to other vehicles, implying it is a more commonly considered vehicles than others. Measuring centrality can be based on various properties of a node, e.g., the number of direct connections to all other nodes (degree), the minimum distance to all other nodes (closeness), and the maximum occurrence on the path of two other nodes (betweenness). **Community** refers to the occurrence of groups of nodes that are more densely connected internally than with the rest of the network [44]. If appropriate communities are detected, the network can be collapsed into a simpler representation without losing much useful information. In a product network, network community analysis [44, 45] helps identify products in the same market segment. As an illustration shown in Fig. 3(a), three distinct communities ("compact sedans" in black, "mid-size sedans" in green, and "SUVs" in red) are identified in a partial vehicle association network. The emergent communities help guide product segmentation and product positioning. Different from centrality and community identified in the unidirectional networks, network **hierarchy** [46] is captured by the directional network based on both consideration and choice data. Hierarchy is formally defined as a strict partially ordered set [47], where each element of the set is a node and the partial ordering (P1<P2) gives an edge from P1 to P2. The directed link reflects customer preference of

one product over the other. In Fig. 3(b), products with high preference ranks (e.g., Mondeo and Malibu) have many incoming links, as indicated by the larger dots.



a. An Illustrative Vehicle Association Network. Nodes are Sized by Network Centralities and Colored by Network Communities.



b. An Illustrative Hierarchical Preference Network. Nodes are Sized by Hierarchy and are Colored as (a).

Figure 3. Centrality, Community, and Hierarchy in Illustrative Networks

In this study, we explain how customer preferences can drive the formation of market structure by analyzing the three metrics in the two vehicle networks. This is done by evaluating the node degree (centrality) and Newman's modularity (community) for Vehicle Association Network and calculating the node in-degree (hierarchy) for the Hierarchical Preference Network, as we detailed next.

5.2. Vehicle Association Network

The vehicle association network is built to help understand what product alternatives are often considered together by a customer. We establish the links between two products to indicate how likely the two products are considered together by a customer in consideration. Using the definition of *lift*, we derive the link strength L_{ij} for product i and j as:

$$lift(i,j) = \frac{\Pr\{\text{co-consider } i \text{ and } j\}}{\Pr\{\text{consider } i\} \cdot \Pr\{\text{consider } j\}},$$

(12)

$$L_{ij} = \begin{cases} lift(i,j), & \text{if } > 1 \\ 0, & \text{otherwise} \end{cases},$$

(13)

A link appears between two vehicles only if the vehicles have positive associations. In the illustrative example in Fig. 3(a), Mondeo is strongly connected to Malibu and Mazda FAW 6, shown by the high link strength. This explains that a customer who considered Mondeo is also very likely to consider Malibu and Mazda FAW 6.

The network centrality is measured by the node degree which counts the total number of links a node has to other nodes. In the vehicle association network, products with a high-degree centrality are those more frequently compared with other vehicles by customers in consideration. As we might expect, most of the high-centrality vehicles are also popular choices in the market, although the two measures are not equivalent.

Nevertheless, the node centrality only looks at the pairs of vehicles that are co-considered by a customer, which is not sufficient to capture all candidate products in customers' considerations. The product community analysis is used instead to determine groups of vehicles that are closely connected. We employ the Newman's optimal modularity method to derive the vehicle community structure. Seven communities are identified and shown in Fig. 4 in different colors. By examining the vehicle models under each community, we may infer the marketing coverage of a brand and other vehicle competitors.

5.3. Hierarchical Preference Network

To further analyze customers' preferences which are beyond considerations, a directional network is constructed using both consideration and purchase data in NCBS. The lift metric shown in Eqn. (16) is slightly modified to accommodate the evaluation of a directional link strength.

$$lift(i \to j) = \frac{\Pr\{\text{co-consider } i \ \&\ j,\ \text{purchase } j\}}{\Pr\{\text{consider } i\} \cdot \Pr\{\text{purchase } j\}},$$

(14)

Again, the links are used to describe positive associations such that a directional link starting from i and point to j has link strength:

$$L_{i \to j} = \begin{cases} lift(i \to j), & \text{if } > 1 \\ 0, & \text{otherwise} \end{cases},$$

(15)

The preference hierarchy between two vehicles of consideration is captured using link directions. For example, in Fig. 3b, a bi-directional (mutual) link can be observed between Mazda FAW 6 and Honda Guangzhou Accord. The existence of mutual link indicates the intense competition between the two cars when both of them are considered and compared, because both cars are purchased by a large

group of customers. However, Mazda FAW6 is slightly more attractive than Honda Grangzhou Accord, because the strength of links in that direction is stronger.

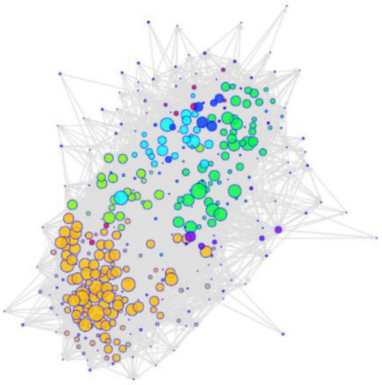


Figure 4. In-degree Hierarchy in Hierarchical Preference Network based on NCBS 2013. Nodes are Sized by Network In-degrees and Colored by Network Communities.

Customers' aggregated preference across the population is reflected by the product ranking generated by the in-degree hierarchy. The in-degree of a node computes the number of incoming links pointed to the node. A node with a high in-degree value implies the corresponding vehicle is very likely to be considered with other vehicles and is also more preferred in customer choice decisions. In NCBS results, we found that some vehicles have been frequently considered (high degree centrality in undirected network), but fall behind in customers' final choices (low in-degree hierarchy).

Based on the illustrations in this section, we conclude that network analysis may serve as a useful tool to determine product priorities in strategic planning. By comparing degree centrality with in-degree hierarchy, we can uncover the root-causes of the differences in vehicle sales under a specific market. These efforts may reveal the specific issues that a marketing team should work on, e.g., social influence and advertisement (low centrality rank), brand market coverage (products not cover certain communities), self-competence (several vehicles in the same community), and product portfolios (low in-degree hierarchy rank), etc.

6. CONCLUSION

In this paper, three different approaches are studied with the goal of gaining marketing insights using customer data. The approach of regional statistics is employed to understand the preference disparities from region to region. As a powerful tool of spatial data mining, geographical maps with bi-polar color progressions relate spatial statistics to customers' attributes and vehicle choices. Next, the impact of various factors are taken into consideration using discrete choice models to describe, explain, and predict customer choice preferences. Preference heterogeneities are treated by introducing customer socio-demographical attributes, random preference coefficients, consideration decisions, subjective ratings as well as brand effects. Furthermore, we employ the network graphs to analyze and visualize the wealth of consideration data that customers generate. Network metrics are applied to covert the complex product association relationships into competitive rankings and market segments for monitoring product positions within a brand or between brand competitors.

We demonstrate the value of the above three approaches using a survey-based data application involving new vehicle sales in China. The complexity of the China auto market's structure and social environment makes it a good benchmark problem to test the validity of these approaches for market-based research. The geographical maps lead us to the possible drivers of customer preferences heterogeneity. The discrete choice models enable quantitative evaluation of these possible effects. The network analysis reveals product associative relations and competitions in the market. The findings from the three approaches may have significant implications for auto manufacturers. For example, they can place their bets on specific market segments and provide optimized product portfolios to take advantage of the on-going preference trends. They can also carry out effective strategic planning to react to the possible market volatility.

It is useful to discuss the advantages and drawbacks of different approaches before we conclude the paper. Regional statistical analysis provides a simplistic way for spatial data description, however; impact factors cannot be quantitatively evaluated, and the result interpretations are judgmental that requires expert knowledge. Unlike regional statistical analysis that examines only one factor at each time, discrete choice analysis allows the analysis of multiple factors and their combined effects in an integrated model. In addition, sensitivity and scenario analysis add extra values to understanding the future market trends and considering the impact of various key drivers. Nevertheless, discrete choice models use trial-and-error in selecting important factors, and limited by specific choice set data structures. Often times, customers' consideration decisions are not well described by discrete choice models. In contrast, network analysis provides insightful information and powerful visualization with respect to product associations derived from customer's consideration decisions. Network analysis is flexible in terms of data structures and can be extended to analyze textual data such as product reviews, customer forums, and industrial news articles. However, customer socio-demographical and usage attributes are not explicitly included in the current network representation we choose which may limit the usage of the approach. In future work, we will explore the use of multidimensional network which integrates customer information into network structures in addition to products' information.

In summary, the three techniques examined in this work play complementary roles in the process of analyzing consumer data. A good understanding of the pros and cons of each approach helps to choose the techniques wisely for achieving a comprehensive understanding of the market and creating useful predictive models.

ACKNOWLEDGMENTS

This research is supported by National Science Foundation (CMMI-1436658). Additional thanks are given to Ford Corporate Economics & Global Consumer Insight groups for providing the China market data.

REFERENCES

1. Wang, Y., Teter J., and Sperling D., *China's soaring vehicle population: Even greater than forecasted?* Energy Policy, 2011. 39(6): p. 3296-3306.
2. Xie, X., Wang Q., and Chen A., *Analysis of competition in Chinese automobile industry based on an opinion and sentiment mining system*. Journal of Intelligence Studies in Business, 2012. 2(1).
3. Abrahams, A.S., et al., *Vehicle defect discovery from social media*. Decision Support Systems, 2012. 54(1): p. 87-97.
4. Murphy, I.B. *How Autometrics Built a Demand Sensing Model with Hundreds of Datasets*. 2013 August 9, 2013; Available from: http://data-informed.com/how-autometrics-built-a-demand-sensing-model-with-hundreds-of-datasets/.
5. Green, P.E. and TuU D.S., *RESEARCH FOR MARKETING, 2/E*. Journal of Marketing, 1970.
6. Pride, W.M., *Marketing decision-making through computer cartography*. Journal of the Academy of Marketing Science, 1977. 5(4): p. 369-378.
7. Train, K.E., *Discrete choice methods with simulation*. 2009: Cambridge university press.
8. Ben-Akiva, M.E. and Lerman S.R., *Discrete choice analysis: theory and application to travel demand*. Vol. 9. 1985: MIT press.
9. Ben-Akiva, M. and Lerman S.R., *Some estimation results of a simultaneous model of auto ownership and mode choice to work*. Transportation, 1974. 3(4): p. 357-376.
10. Lerman, S.R. and Ben-Akiva M., *Disaggregate Behavior Model of Automobile Ownership*. Transportation Research Record, 1976(569).
11. Lave, C.A. and Train K., *A disaggregate model of auto-type choice*. Transportation research part A: general, 1979. 13(1): p. 1-9.
12. Boyd, J.H. and Mellman R.E., *The effect of fuel economy standards on the US automotive market: an hedonic demand analysis*. Transportation Research Part A: General, 1980. 14(5): p. 367-378.
13. Berry, S., Levinsohn J., and Pakes A., *Automobile prices in market equilibrium*. Econometrica: Journal of the Econometric Society, 1995: p. 841-890.
14. Brownstone, D., Bunch D.S., and Train K., *Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles*. Transportation Research Part B: Methodological, 2000. 34(5): p. 315-338.
15. Brownstone, D. and Train K., *Forecasting new product penetration with flexible substitution patterns*. Journal of econometrics, 1998. 89(1): p. 109-129.
16. Train, K.E. and Winston C., *Vehicle Choice Behavior and Declining Market Share of US Automakers*. International Economic Review, 2007. 48(4): p. 1469-1496.
17. Bunch, D.S., et al., *Demand for clean-fuel vehicles in California: a discrete-choice stated preference pilot project*. Transportation Research Part A: Policy and Practice, 1993. 27(3): p. 237-253.
18. Wassenaar, H.J. and Chen W., *An approach to decision-based design with discrete choice analysis for demand modeling*. Journal of Mechanical Design, 2003. 125(3): p. 490-497.
19. Wassenaar, H.J., et al., *Enhancing discrete choice demand modeling for decision-based design*. Journal of Mechanical Design, 2005. 127(4): p. 514-523.
20. Shiau, C.-S.N. and Michalek J.J., *Optimal product design under price competition*. Journal of Mechanical Design, 2009. 131(7): p. 071003.
21. Kumar, D., Chen W., and Simpson T.W., *A market-driven approach to product family design*. International Journal of Production Research, 2009. 47(1): p. 71-104.
22. Orsborn, S., Cagan J., and Boatwright P., *Quantifying aesthetic form preference in a utility function*. Journal of Mechanical Design, 2009. 131(6): p. 061001.
23. He, L., et al., *Choice modeling for usage context-based design*. Journal of Mechanical Design, 2012. 134(3): p. 031007.
24. He, L., Hoyle C., and Chen W., *Examination of customer satisfaction surveys in choice modelling to support engineering design*. Journal of Engineering Design, 2011. 22(10): p. 669-687.
25. Hoyle, C., et al., *Integrated Bayesian hierarchical choice modeling to capture heterogeneous consumer preferences in engineering design*. Journal of Mechanical Design, 2010. 132(12): p. 121010.
26. Tversky, A. and Kahneman D., *The framing of decisions and the psychology of choice*. Science, 1981. 211(4481): p. 453-458.
27. Shocker, A.D., et al., *Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions*. Marketing letters, 1991. 2(3): p. 181-197.
28. Netzer, O., et al., *Mine your own business: Market-structure surveillance through text mining*. Marketing Science, 2012. 31(3): p. 521-543.
29. Pandey, G., et al., *Association rules network: Definition and applications*. Statistical analysis and data mining, 2009. 1(4): p. 260-279.
30. Raeder, T. and Chawla N.V., *Market basket analysis with networks*. Social network analysis and mining, 2011. 1(2): p. 97-113.
31. Parsons, G.R. and Kealy M.J., *Randomly drawn opportunity sets in a random utility model of lake recreation*. Land Economics, 1992: p. 93-106.
32. Luce, R.D. and Suppes P., *Preference, utility, and subjective probability*. 1965: Wiley.
33. Berry, S., Levinsohn J., and Pakes A., *Differentiated products demand systems from a combination of micro and macro data: The new car market*. 1998, National bureau of economic research.
34. Peterson, R.A. and Wilson W.R., *Measuring customer satisfaction: fact and artifact*. Journal of the Academy of Marketing science, 1992. 20(1): p. 61-71.
35. Rossi, P.E., Gilula Z., and Allenby G.M., *Overcoming scale usage heterogeneity: A Bayesian hierarchical approach*. Journal of the American Statistical Association, 2001. 96(453): p. 20-31.
36. Guadagni, P.M. and Little J.D., *A logit model of brand choice calibrated on scanner data*. Marketing science, 1983. 2(3): p. 203-238.
37. Winer, R.S., *A reference price model of brand choice for frequently purchased products*. Journal of consumer research, 1986: p. 250-256.
38. Klaiber, H.A. and von Haefen R.H., *Incorporating random coefficients and alternative specific constants into discrete choice models: implications for in-sample fit and welfare estimates*. Western Regional Research, 2008: p. 200.
39. Tan, P.-N., Kumar V., and Srivastava J., *Selecting the right objective measure for association analysis*. Information Systems, 2004. 29(4): p. 293-313.
40. Real, R. and Vargas J.M., *The probabilistic basis of Jaccard's index of similarity*. Systematic biology, 1996: p. 380-385.
41. Chowdhury, G., *Introduction to modern information retrieval*. 2010: Facet publishing.
42. Freeman, L.C., *Centrality in social networks conceptual clarification*. Social networks, 1979. 1(3): p. 215-239.
43. Wasserman, S., *Social network analysis: Methods and applications*. Vol. 8. 1994: Cambridge university press.
44. Newman, M.E. and Girvan M., *Finding and evaluating community structure in networks*. Physical review E, 2004. 69(2): p. 026113.
45. Clauset, A., Newman M.E., and Moore C., *Finding community structure in very large networks*. Physical review E, 2004. 70(6): p. 066111.
46. De Vries, H., *Finding a dominance order most consistent with a linear hierarchy: a new procedure and review*. Animal Behaviour, 1998. 55(4): p. 827-843.
47. Corominas-Murtra, B., et al., *On the origins of hierarchy in complex networks*. Proceedings of the National Academy of Sciences, 2013. 110(33): p. 13316-13321.