August 2017

Specification of Mixed Logit Models Using an Optimization Approach

Cristian David Arteaga Sanchez
University of Nevada, Las Vegas, cristiandavidarteaga@gmail.com

SPECIFICATION OF MIXED LOGIT MODELS USING AN OPTIMIZATION APPROACH

By

Cristian David Arteaga Sanchez

Bachelor in Computer Science
College of Engineering
Universidad del Cauca, Colombia
2015

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science in Engineering – Civil and Environmental Engineering

Department of Civil and Environmental Engineering and Construction
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
August 2017

UNLV | GRADUATE COLLEGE

Thesis Approval

The Graduate College
The University of Nevada, Las Vegas

August 23, 2017

This thesis prepared by

Cristian David Arteaga Sanchez

entitled

Specification of Mixed Logit Models Using an Optimization Approach

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Engineering – Civil and Environmental Engineering
Department of Civil and Environmental Engineering and Construction

Alexander Paz, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Interim Dean

Mohamed Kaseko, Ph.D.
Examination Committee Member

Dave James, Ph.D.
Examination Committee Member

Brendan Morris, Ph.D.
Examination Committee Member

Justin Zhan, Ph.D.
Graduate College Faculty Representative

ABSTRACT

Mixed logit models are a widely-used tool for studying discrete outcome problems. Modeling development entails answering three important questions that highly affect the quality of the specification: (i) what variables are considered in the analysis? (ii) what are going to be the coefficients for these variables? and (iii) what density function these coefficients will follow? The literature provides guidance; however, a strong statistical background and an ad hoc search process are required to obtain the best model specification. Knowledge of the problem context and data is required. Given a dataset including discrete outcomes and associated characteristics the problem to be addressed in this thesis is to investigate to what extend a relatively simple metaheuristic such as Simulated Annealing, can determine the best model specification for a mixed logit model and answer the above questions. A mathematical programing formulation is proposed and simulated annealing is implemented to find solutions for the proposed formulation. Three experiments were performed to test the effectiveness of the proposed algorithm. A comparison with existing model specifications for the same datasets was performed. The results suggest that the proposed algorithm is able to find an adequate model specification in terms of goodness of fit thereby reducing involvement of the analyst.

ACKNOWLEDGMENT

I wish to express my deep gratitude to my advisor Dr. Alexander Paz for his mentoring and encouragement throughout this research project. Also, for all his advices and supportive words for the development of my professional career. I want to thank the members of my committee- Dr. Mohamed Kaseko, Dr. Dave James, Dr. Justin Zhan, and Dr. Brendan Morris- for their commitment and dedication to teaching excellence. I want to thank my friends and colleagues Carlos Gaviria, Victor Molano, Mayra Sarria, Kul Shresta, and Daniel Emaasit for their support and friendship. Finally, I would like to thank Ms. Julie Longo for all her teachings and help.

DEDICATION

To my family, I cannot thank enough all what they have done for me. To my wife Nathali, for her unconditional support, dedication and love, thank you for always being there. To God, who makes everything possible and all dreams come true.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGMENT iv

DEDICATION v

TABLE OF CONTENTS vi

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: LITERATURE REVIEW 6

CHAPTER 3: METHODOLOGY 9

Mathematical Programming - Problem Formulation 9

Solution Algorithm 11

CHAPTER 4: EXPERIMENTS 15

Experiment 1 and 2 15

Experiment 3 17

CHAPTER 5: RESULTS 18

Experiment 1 and 2 18

Experiment 3 22

CHAPTER 6: CONCLUSIONS 26

APPENDIX A: INSTRUCTIONS TO EXECUTE ALGORITHM 29

Requirements 29

Steps 29

APPENDIX B: SOURCE CODE FOR ALGORITHM 30
Files Structure 30
Source Code 31
mxlogit_search.R 31
mxlogit_search_fun.R 33
params.R 38
REFERENCES 40
CURRICULUM VITAE 45

LIST OF TABLES

Table 1. Variables for Alternative-Fuel Vehicles Dataset..... 16

Table 2. Variables for Streaming Video Service Dataset 17

Table 3. Algorithm output for Experiments 1 and 2. 20

Table 4. Summary of Quality Measures for Models..... 22

Table 5. Algorithm output for Experiment 3..... 24

LIST OF FIGURES

Figure 1. Steps of the proposed simulated annealing algorithm. 14

Figure 2. BIC vs. iterations for Model 1a. 18

Figure 3. BIC vs. iterations for Model 1b. 19

Figure 4. BIC vs iterations for Model 2. 23

CHAPTER 1: INTRODUCTION

Modeling and prediction of discrete outcomes is a common problem in many areas, including among others economics, engineering, and medicine. Some examples of discrete outcome problems include: (i) analysis of transportation modes (i.e., car, transit, or walking) based on observed socioeconomic characteristics, (ii) estimate the presence of a pathology based on attributes of a patient, and (iii) estimate how many cars will be owned based on observed characteristics of a household.

In general, a categorical variable associated or explained by a set of attributes and/or characteristics can be considered a discrete outcome problem (Train, 2003). In transportation, discrete outcome analysis has a wide range of applications. In land use modeling, it is applied for choices of residential locations based on observed demographic attributes of people and characteristics of the locations (Wegener, 2004). In route choice analysis, discrete outcome models are used for prediction of route choices, based on observed attributes of both travelers and available routes (Paz, Emaasit, & de la Fuente, 2016; Paz & Peeta, 2009) . In traffic safety, prediction of crash severity based on roadway characteristics, driver behavior and weather factors (Milton, Shankar, & Mannering, 2008). In travel demand analysis, choices for auto and bike ownership based on attributes of travelers (Pinjari, Pendyala, Bhat, & Waddell, 2011).

Several statistical and machine-learning approaches have been proposed in the literature to model discrete outcome problems (Luo, 2015; Omrani, 2015). In the machine learning side, techniques such as artificial neural networks and support vector machines have been successfully applied. In statistics, models such as logit, probit, nested logit, mixed logit have been extensively used (Train, 2003). Machine learning has showed superior predictive ability compared to statistical models (Karlaftis & Vlahogianni, 2011). However, one disadvantage of machine learning

approaches is that these are considered 'black box' methods. These approaches, although useful for prediction, do not provide additional insights about the data. In the other hand, statistical techniques have a significant advantage in terms of interpretation. The output of statistical models is a set of coefficients whose values are intuitive and have a meaningful interpretation. Also, statistical models can derive useful measures such as marginal effects, elasticities, willingness to pay, among others (Hensher & Ton, 2000).

Regardless of the proposed approach, the researcher needs to decide which variables are to be considered in the model specification. The modeling process is time consuming and subject to expert knowledge and ad hoc trial and error approaches. The variables included in a model highly affect its predictive performance. Models with a proper and smallest subset of explanatory variables allow larger influence of the included variables, eliminate redundancy, provide a better understanding of the final model, reduce costs of data acquisition and are computationally efficient (Fouskakis & Draper, 2008). Variable selection, also referred in the literature as subset selection or model specification, aims to find a model with the highest explanatory power while selecting the smallest possible number of variables. A challenge is that the number of possible combinations of variables that could be considered grows exponentially as the number of potential explanatory variables increases (Sato, Takano, Miyashiro, & Yoshise, 2016; Vinterbo & Ohno-Machado, 1999). For example, for a model with 30 variables the number of different possible specifications is 2^{30}= 1,073,741,824. This is computationally intensive to be solved using an exhaustive search. Various approaches used to address this problem are described below in the literature review.

Discrete outcome problems can be viewed as discrete choice processes where a decision maker chooses an alternative from a finite set. Theoretically, it is assumed that the chosen alternative maximizes the utility of the decision maker. This is known as random utility

maximization. Random because of the inability to observe all the factors that impact the utility. This means that the utility is calculated using observed factors and making assumptions about the distribution of unobserved factors, also known as error terms. (Ben-Akiva & Lerman, 1985; Train, 2003).

Multinomial logit and probit are common choice models that have been successfully applied for modeling discrete outcome problems. It is known that logit models suffer limitations such as Independence of Irrelevant Alternatives (IIA), restrictive substitution patterns and inability to model random taste variation. Probit models have addressed these limitations; however, they are restricted to model random taste variation using only the normal distribution, which is not always convenient. In view of these limitations, Mixed logit models have been proposed (Train, 2003) as one of the most prominent techniques for modeling discrete outcome problems.

Mixed logit models address the limitations of logit and probit by allowing modeling of variables with random coefficients. Such variables can follow any statistical distribution specified by the researcher, and a general random term that follows an extreme value distribution. The predictive power and quality of a mixed logit highly depends on an appropriate definition of the distribution of the random coefficients (Hensher & Greene, 2003). The modeling of coefficients as random variables provided by mixed logit allows to capture heterogeneity in preferences among the decision makers. For example, in a mixed logit model for vehicle choices, a variable such as fuel consumption modeled as random and normally distributed, with a mean value of -0.3 and a standard deviation 1.2 can be understood as: given that the mean is slightly below zero, people have more inclination for cars with lower fuel consumption, however the standard deviation evidences that a significant portion of people are willing to have a car with higher fuel consumption.

Modeling with random coefficients is not the only type of derivation for mixed logit models.

Another widely applied derivation is the use of error components to model correlations between the utilities for the alternatives. The choose of what type of derivation for mixed logit to use depends entirely on the needs of the analyst. When the purpose is to analyze the heterogeneity in preferences, then the derivation of mixed logit with random terms is more suitable. In the other hand, if the analyst needs to study the different correlation patterns generated by the error terms, then the derivation with error terms fits better this context. The derivations of mixed logit as random terms or error components are equivalent with the only difference being the interpretation (Train, 2003). For this study, the derivation of random terms for mixed logit was used.

The output of a mixed logit with random coefficients includes the mean and standard deviation of the variables treated as random terms. The mean represents the average preference about the variable while the standard deviation has valuable information about the heterogeneity of that preference , in other words how dispersed is the preference (Daniel McFadden and Kenneth Train, 2000). For example, in a mixed logit model for vehicle choices, a variable such as fuel consumption modeled as random and normally distributed, with a mean value of -0.3 and a standard deviation 1.2 can be understood as: given that the mean is slightly below zero, people have more inclination for cars with lower fuel consumption, however the standard deviation evidences that a significant portion of people are willing to have a car with higher fuel consumption.

Given a mixed logit estimation problem, several assumptions are required to determine the best model specification. In general, the distribution of the random coefficients, and potential explanatory variables need to be assumed before a model is estimated (Hensher & Greene, 2003). This study, proposes an optimization framework to search the best model specification including the variables to be considered, the coefficients as well as the distribution and associated parameters for the corresponding coefficients. In addition, a solution algorithm was implemented and tested

with two datasets.

CHAPTER 2: LITERATURE REVIEW

Variable selection is a topic of high interest in the scientific community. Substantial intellectual effort has been invested in characterizing and solving this problem since the early 60's (Efromyson, 1960). Interest on this problem grows, as new modeling techniques appear, and the availability of data increases with new advances in technology. For any statistical model, when all the possible explanatory variables are included, several issues can arise. For example, irrelevant variables may suppress important relationships between other variables or correlated variables create multicollinearity. A balance is recommended with a number of variables not too small or too large (Hasan Örkcü, 2013) while providing adequate predictive performance (Kadane & Lazar, 2004).

Variable selection approaches have been classified as filter, wrapper and embedded methods based on the strategy used to search a subset of variables (Mehmood, Liland, Snipen, & Sæbø, 2012). Branch and bound algorithms along with stepwise variable inclusion/elimination are common wrapper variable selection methods. These methods have proven to be effective in subset selection for partial less squared regression and principal component analysis as well as logistic regression. A disadvantage of the stepwise approach is that its performance decreases for problems with a number of variables greater than 30 (Brusco, 2014).

To perform variable selection, it is required to have a quality measure to quantify how good a model specification is. In other words, a measure that allows to compare models (Kadane & Lazar, 2004). Several approaches have been used for this purpose. Bayesian Information Criteria (BIC) also known as Swartz Information Criteria has been successfully employed for model comparison in variable selection for continuous and discrete outcome problems (Sato et al., 2016). This measure initially proposed by Schwarz (1978) has been applied in several variable selection problems (Sato et al., 2016; Tutz, Pößnecker, & Uhlmann, 2015; Vicari & Alfó, 2014). BIC uses

the likelihood as goodness of fit measure and it includes a penalization term for the number of parameters used to obtain such likelihood. BIC is similar to Akaike Information Criterion, which is also used for models comparison; however, BIC provides larger penalization for the number of parameters. Prediction accuracy, which measures the percentage of outcomes correctly classified, has been used in discrete outcome problems (Brusco & Steinley, 2011). The Wilks' lambda measure has been applied for similar problems in principal component analysis (Pacheco, Casado, & Porras, 2013).

Simulated annealing is a metaheuristic extensively used to solve optimization problems (Kirkpatrick, Gelatt, & Vecchi, 1983). This metaheuristic has been applied in variable selection problems (Lin, Lee, Chen, & Tseng, 2007; Meiri & Zahavi, 2004; Sutter & Kalivas, 1993). It has proven to outperform other methods including stepwise elimination and branch and bound. The main challenge of simulated annealing is the need to define algorithm parameters. The performance of simulated annealing highly depends of proper specification of its parameters (Brusco, 2014).

Variable selection approaches for logit and probit models using optimization metaheuristics has been successfully performed (Changpetch & Lin, 2013; Fouskakis & Draper, 2008; Pacheco, Casado, & Núñez, 2009; Sato et al., 2016; Vinterbo & Ohno-Machado, 1999; Zahid & Tutz, 2013). Tabu search algorithm has been used for variable selection in logistic regression outperforming forward and backward elimination (Pacheco et al., 2009). Fousakis and Draper, (2008) performed a comparison of heuristic optimization methods for selection of binary-outcome logit models. Additional to variable selection, the optimization algorithm included a budget constraint component. Association rules analysis for selection of multinomial logit has been proposed as a novel method to identify variable interactions. (Changpet & Lin, 2013). Additionally, mixed

integer optimization with a piecewise approximation of the logistic loss function has been applied for variable selection in logistic regression (Sato et al., 2016).

The search of a mixed logit specification is more involved compare to logit or probit because the algorithms must determine what coefficients are deterministic or stochastic as well as the corresponding distributions. To determine these configurations for a mixed logit model, the literature provides guidance. Train (2003) provides the theoretical background necessary for the estimation and interpretation of mixed logit models. The adequacy of coefficients modeled as random parameters can be determined with a test of omitted variable and properly defined artificial variables (Daniel McFadden and Kenneth Train, 2000). Marginal likelihood with Bayesian approaches has been proposed as a comparison measure for mixed logit.(Balcombe, Chalak, & Fraser, 2009). For modeling of correlation and account ford scale heterogeneity Hess & Train (2017) provide a list of suggestions that the analyst can or should use to approach this specification. To the best knowledge of the authors, an approach to search the best mixed logit model specification is not yet available in the literature. For the remaining of this document, a mixed logit model specification will be known as *model specification*.

CHAPTER 3: METHODOLOGY

Mathematical Programming - Problem Formulation

The following notation is used to describe and formulate the proposed problem:

$\boldsymbol{x}$	vector of potential explanatory variables
N	number of observations
K	number of potential explanatory variables
S	number of included variables
J	number of alternatives or discrete outcomes
i	subscript to denote a decision maker; $i = 1, 2, \ldots, N$
j	superscript to denote an alternative; $j = 1, 2, \ldots, J$
k	subscript for a variable, $k = 1, 2, ..., K$
y_{ij}	indicator variable equal to 1 if decision maker i chooses alternative j; 0 otherwise.
s_k	indicator variable to denote when variable x_k is included, $s_k \in \boldsymbol{s}$. s_k equal to 1 if variable x_k is included; 0 otherwise.
β_k^j	coefficient for variable x_k and alternative j; $\beta_k^j \in \boldsymbol{\beta}$.
$\boldsymbol{s}$	vector of included variables.
$\boldsymbol{\beta}$	vector of coefficients for potential explanatory variables.
$\boldsymbol{f}$	vector of density functions for coefficients $\boldsymbol{\beta}$.
f_k	density function for coefficient β_k. Possible density functions f_k are: normal, lognormal, uniform, triangular or f_k is equatl to when no density function will be used

The observed utility V_{ij} that a decision maker i obtains from alternative j can be represented as a linear dependency on the attributes of the decision maker and the alternatives as:

$$V_{ij} = \beta_0^j + \beta_1^j x_{i1} + \cdots + \beta_K^j x_{iK} \quad (1)$$

For this research, the observed portion of utility V_{ij} is extended to add the indicator s_k of included variables.

$$V_{ij} = \beta_0^j + \beta_1^j x_{i1} s_1 + \cdots + \beta_K^j x_{iK} s_K \quad (2)$$

In mixed logit, the probability that a decision maker *i* chooses alternative *l* is modeled as (Train, 2003):

$$P_{il} = \int \frac{e^{V_{il}}}{\sum_{j=1}^{J} e^{V_{ij}}} \boldsymbol{f}(\boldsymbol{\beta}) d\boldsymbol{\beta} \quad (3)$$

The coefficients $\boldsymbol{\beta}$ can be estimated by maximum log- likelihood estimation (MLE). The log-likelihood LL, is calculated as:

$$LL = ln(L) = \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} ln(P_{ij}) \quad (4)$$

BIC, which is the measure for model comparison used in this study, is represented by Equation (5).

$$BIC = \ln(N)\, S - 2\ln(LL) \quad (5)$$

The objective, represented by Equation (6), is to find the model specification $\boldsymbol{M} = \{\boldsymbol{s}, \boldsymbol{f}\}$ with included variables $\boldsymbol{s}$, the coefficients $\boldsymbol{\beta}$, and the density functions $\boldsymbol{f}$ that maximize the BIC.

$$Min\ BIC = \ln(N)\, S$$

$$- 2\ln\Big(\sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} ln\Big(\int \frac{e^{\beta_0^j + \beta_1^j x_{i1} s_1 + \cdots + \beta_K^j x_{iK} s_K}}{\sum_{j=1}^{J} e^{\beta_0^j + \beta_1^j x_{i1} s_1 + \cdots + \beta_K^j x_{iK} s_K}} \boldsymbol{f}(\boldsymbol{\beta}) d\boldsymbol{\beta} \Big) \Big) \quad (6)$$

Subject to:

$$s_k = \begin{Bmatrix} 1 & \leftrightarrow variable\ x_k\ is\ included\ in\ the\ model\ for\ all\ k = 1, 2, \dots N; \\ 0 & otherwhise \end{Bmatrix} \quad (6)$$

Solution Algorithm

A simulated annealing algorithm was used to solve the above minimization problem. This metaheuristic was selected because it has been successfully applied in variable selection problems (Brusco, 2014; Hasan Örkcü, 2013). In addition, its implementation and parameter tuning are relatively easy. Simulated annealing is a widely-used metaheuristic for optimization problems (Kirkpatrick et al., 1983) which uses the analogy of the controlled cooling process of materials to improve their properties (annealing process).

Simulated annealing iteratively searches the feasible region trying to find better solutions. One of the most important features of simulated annealing is that it avoids local optimal by strategically accepting bad quality solutions. The probability of accepting a bad solution is a function of the temperature. At the beginning of the optimization process, when the temperature is high, the algorithm accepts low quality solutions with a high probability. The acceptance probability decreases as the temperature value decreases.

To use a simulated annealing algorithm a researcher needs to specify: (i) a quality measure for a solution (BIC in this case), (ii) a neighborhood criteria that tells the algorithm how to move through the search space and (iii) a cooling schedule (Initial temperature T_0, final minimum temperature T_{min}, cooling rate ϕ, and Boltzmann Constant B) that models how the temperature decreases and when the algorithm stops. The stopping criteria for the algorithm is also handled by the cooling schedule, specifically by the minimum temperature. The cooling schedule for the algorithm proposed in this study was configured to execute 150 iterations.

The algorithm steps are illustrated in Figure 1 and a description of such steps is provided below.

Step 1: Initialization

Step 1.1: An initial solution $\boldsymbol{M} = \{\boldsymbol{s}, \boldsymbol{f}\}$ is generated by randomly assigning values to $\boldsymbol{s}$ and $\boldsymbol{f}$.

Step 1.2: Set values of initial temperature (T_0), minimum temperature (T_{min}), cooling rate (ϕ), and the maximum number of neighbors to be generated (N_{max}) at each temperature level.

Step 1.3: Initialize value for current temperature T as $T = T_0$

Step 2: Generate neighbor solution M_n

Step 2.1: A neighbor solution is generated from $\boldsymbol{M} = \{\boldsymbol{s}, \boldsymbol{f}\}$ by randomly changing one element in the vector of selected variables and in the vector of density functions.

Step 2.2: Step 2.1 is repeated until N_{max} neighbor solutions have been generated.

Step 2.3: For each N_{max} neighbor, estimate mixed logit model and remove not significant variables at 0.1 level.

Step 2.4: Calculate BIC for all N_{max} neighbors generated in Step 2.3. Only the variables that were established as significant in previous step are used for the estimation of the BIC measure.

Step 2.5: Select the best, smaller BIC, quality solution M_n from the N_{max} neighbor solutions.

Step 3: Determine acceptance of neighbor solution M_n

Step 3.1: If the neighbor solution M_n has a BIC smaller than current solution M then M_n is set as current solution M, ($M = M_n$). Otherwise go to step 3.2.

Step 3.2: Generate a random number $r = R\ (0,1)$.

Step 3.3: Calculate $\Delta BIC = BIC(M_n)\text{-}BIC(M)$.

Step 3.4: Calculate the probability of acceptance $Pa = exp(\Delta BIC/B*T)$

Step 3.5: If $Pa > r$ then Mn is set as current solution M, ($M = M_n$).

Step 4: Check stop criteria.

Step 4.1: If $T < T_{min}$ (the cooling was completed) then stop and return the current solution M. Otherwise go to step 4.2

Step 4.2: Update the temperature $T = \phi T$ and return to step 2.

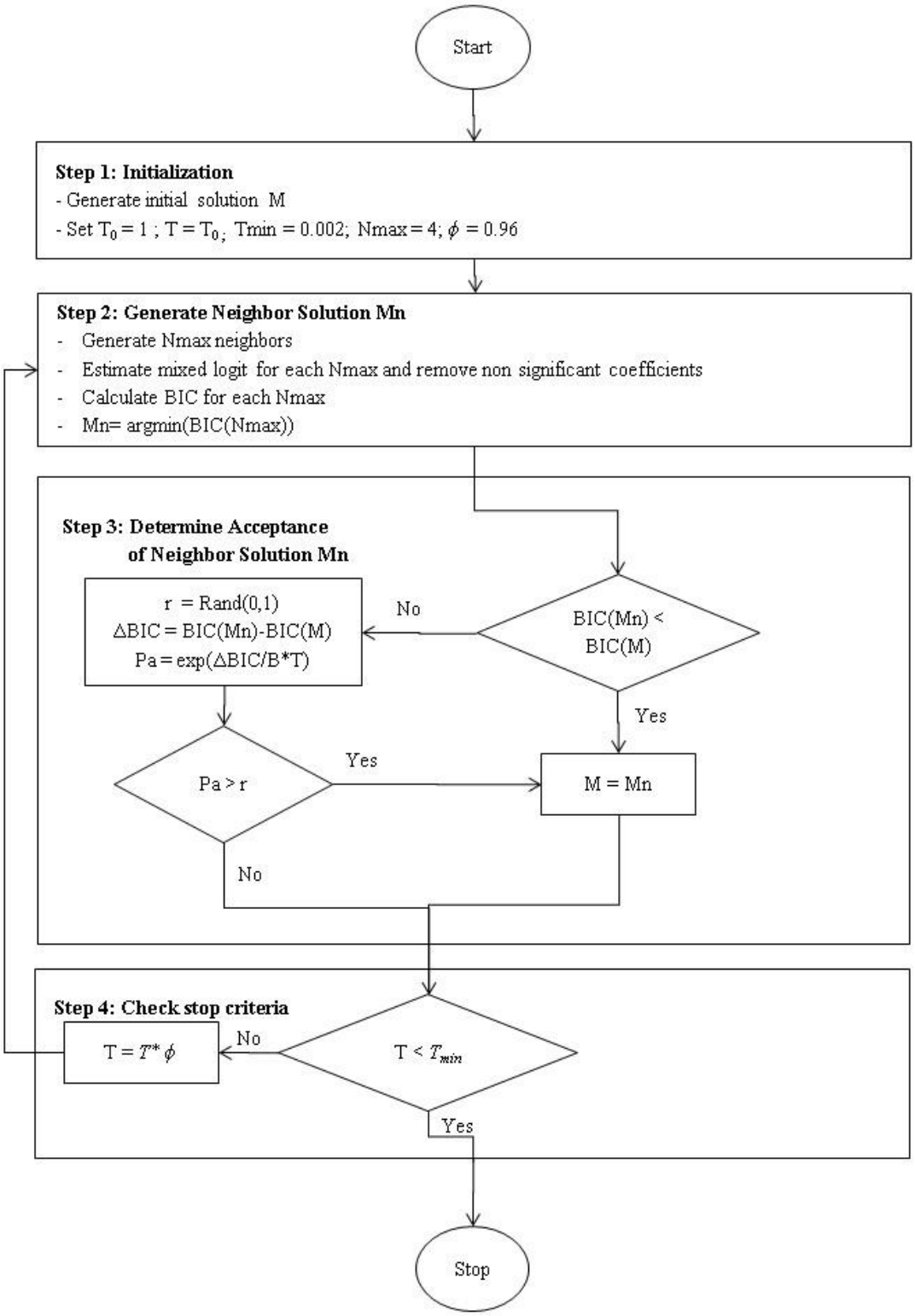

Figure 1. Steps of the proposed simulated annealing algorithm.

CHAPTER 4: EXPERIMENTS

Experiment 1 and 2

In this study, three experiments were performed. For the first and second experiment a dataset for choices of alternative-fuel vehicles, initially used by Brownstone & Train (2000), was used. This dataset comes from a stated preference survey with 21 alternative specific variables and 4,654 observed choices. Table 1 provides a description of the variables included on this dataset as shown in Brownstone & Train (1999).

The first experiment had a random start point and parameters for simulated annealing: T_0 = 1, minimum temperature T_{min} = 0.002, cooling rate ϕ = 0.96, and Boltzmann constant = 0.0009. The neighboring generation process changes 15% of the elements in the vector of the selected variables and in the vector of density functions. The output of this experiment is denoted as Model 1a.

The second experiment has a different start point and a slight modification of the parameters of simulated annealing. A specification similar to the one in McFadden & Train (Daniel McFadden and Kenneth Train, 2000) was used as start point. Hence, the starting search point is already an excellent solution. The motivation of this experiment is to represent a more extensive search relative to the first experiment and to investigate the existence of a better model specification subject to the use of a superior optimization algorithm. In the context of metaheuristic optimization, an intensive search involves both ‘exploration’ and ‘exploitation’. The Boltzmann constant was set to 0.03; for a neighboring generation, and only one element in the vector of selected variables and vector of density functions was changed. The cooling schedule and neighborhood criteria were set to perform a more intensive search. The output of this experiment is denoted as Model 1b.

Table 1. Variables for Alternative-Fuel Vehicles Dataset

Variable names	Description
Price/ln(income)	Purchase price in thousands of dollars, divided by the natural log of household income in thousands
Range	Hundreds of miles that the vehicle can travel between refueling/recharging
Acceleration	Seconds required to reach 30 mph from stop, in tens of seconds (e.g., 3 s is entered as 0.3)
Top Speed	Highest speed that the vehicle can attain, in hundreds of miles/h (e.g., 80 mph is entered as 0.80)
Pollution	Tailpipe emissions as fraction of comparable new gas vehicle
Size	0"mini, 0.1"subcompact, 0.2"compact, 0.3"mid-size or large
Big Enough	1 if household size is over 2 and vehicle size is 3; 0 otherwise
Luggage Space	Luggage space as fraction of comparable new gas vehicle
Operating Cost	Cost per mile of travel, in tens of cents per mile (e.g., 5 cents/miles is entered as 0.5.) For electric vehicles, cost is for home recharging. For other vehicles, cost is for station refueling
Station Availability	Fraction of stations that have capability to refuel/recharge the vehicle
Sports Utility Vehicle	1 for sports utility vehicle, zero otherwise
Sports Car	1 for sports car, zero otherwise
Station Wagon	1 for station wagon, zero otherwise
Truck	1 for truck, zero otherwise
Van	1 for van, zero otherwise
EV	1 for electric vehicle, zero otherwise
Commute <5 & EV	1 if respondent commutes less than five miles each day and vehicle is electric; zero otherwise
College & EV	1 if respondent had some college education and vehicle is electric; zero otherwise
CNG	1 for compressed natural gas vehicle, zero otherwise
Methanol	1 for methanol vehicle, zero otherwise
College & methanol	1 if respondent had some college education and vehicle is methanol; zero otherwise

Experiment 3

For the third experiment, a dataset from a stated preference survey for streaming video services was used. This dataset with 9 variables and 3300 observations was initially used by Glasgow & Butler (2017). This dataset has responses for 330 individuals; each individual has 10 observed choices which constitutes it as panel data. Table 2 provides a description of the variables included on this dataset as shown by Glasgow & Butler (2017). The third experiment has the same parameters of first experiment for simulated annealing algorithm; the only difference being the Boltzmann constant which is 0.0004 in this case.

Table 2. Variables for Streaming Video Service Dataset

Variable	Description
Share NPII	1 for Share Non-Personally Identifiable Information, zero otherwise
Share NPII and PII	1 for Share Non-Personally Identifiable Information and Personally Identifiable Information, zero otherwise
Price	Monthly price of the service
More content	1 for 10 000 movies, 5000 TV episodes, zero otherwise
More TV/fewer movies	1 for 2000 movies, 13 000 TV episodes, zero otherwise
Commercials	1 for Commercials, 0 otherwise, zero otherwise
Fast content	1 for TV episodes next day, movies in 3 months, zero otherwise
No service	1 for no streaming video service, zero otherwise

For all the experiments, the parameters for simulated annealing were defined by following suggestions from previous studies by Hajek (Hajek, 1988), Nourani & Andresen (Nourani & Andresen, 1998) and Paz et. al. (Paz, Molano, Martinez, Gaviria, & Arteaga, 2015). R programming language was used for the implementation of the proposed algorithm, and open source library, mlogit for R, was used to estimate of the mixed logit models (Croissant, 2012). Halton sequences with 100 random draws were used for the estimation. The experiments were executed on a laptop with 6 GB of RAM memory and an i7-4500U processor at 1.8 GHz.

CHAPTER 5: RESULTS

Experiment 1 and 2

Figures 2 and 3 illustrate the improvement of the BIC over iterations of the proposed algorithm for Experiments 1 and 2, respectively. For Experiment 1, the initial BIC was 15,749.4; after 150 iterations, the BIC was 14,946.33. The execution time was 13.4 hours. As illustrated in Figure 3, similar results were obtained for Experiment 2. These improvements in the BIC suggest that the proposed algorithm can find a model specification with adequate goodness of fit. Table 3, provides the output of the proposed algorithm for Experiments 1 and 2; these are the models with the minimum BICs.

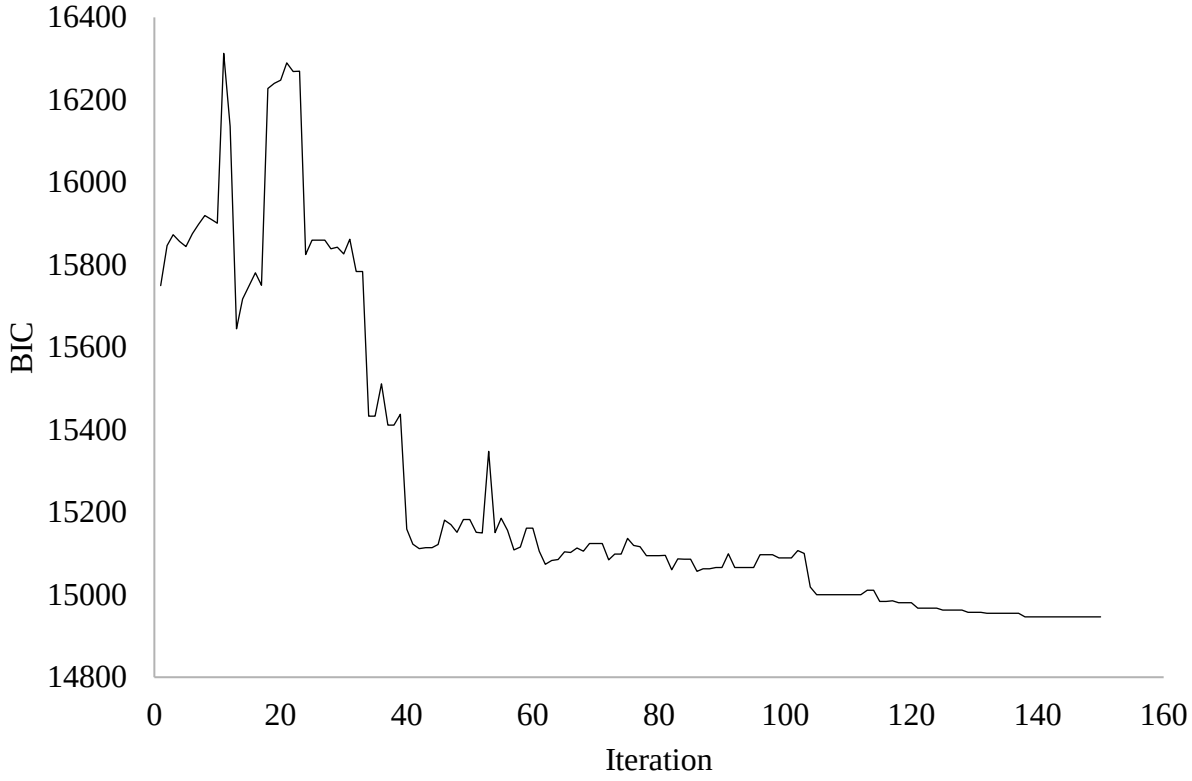

Figure 2. BIC vs. iterations for Model 1a.

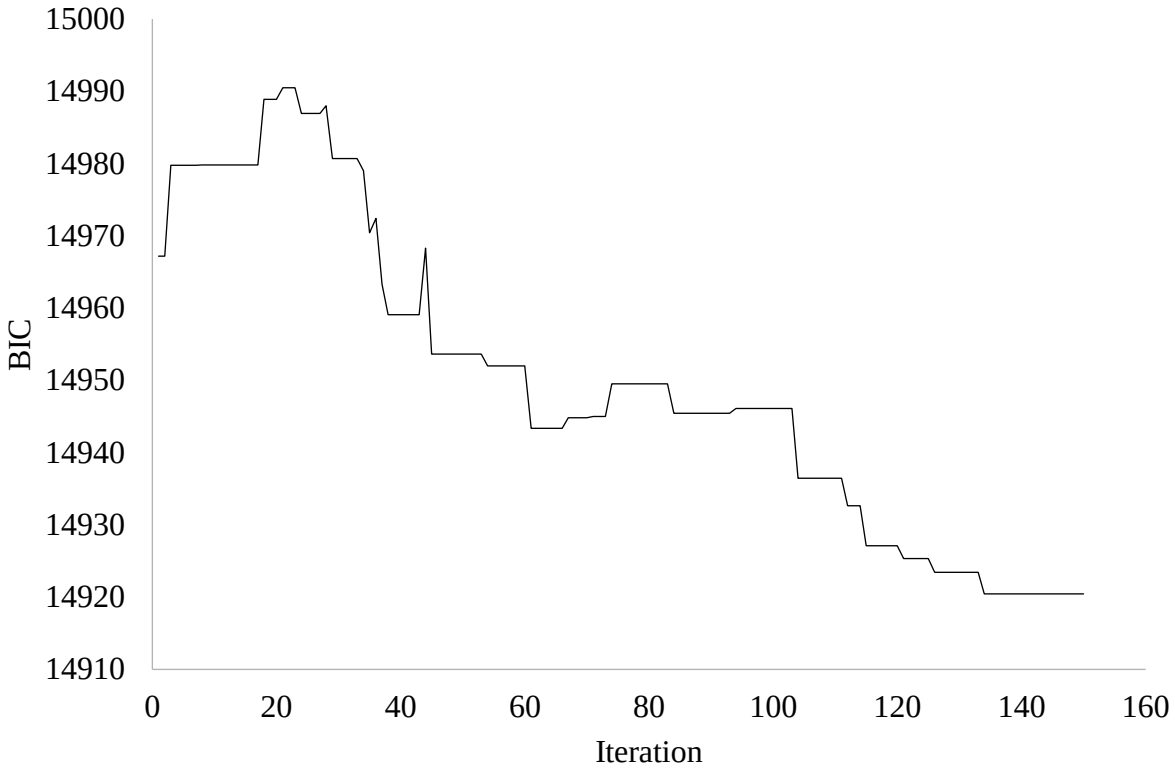

Figure 3. BIC vs. iterations for Model 1b.

For Model 1a, the random effects of variables Acceleration and Operating Cost follow a normal distribution. For Model 1b, the random effects of variables Size and EV follow a normal distribution while for variables Operating Cost and CNG follow a triangular distribution. The use of triangular distributions has benefit when calculating the willingness to pay values. McFadden & Train (Daniel McFadden and Kenneth Train, 2000) estimated (Table IV) a mixed logit model for the same dataset used in this study. This model from McFadden & Train is denoted here as MAT.

Table 3. Algorithm output for Experiments 1 and 2.

	Model 1a		Model 1b	
Variable	**Coefficient**	**Std. Error**	**Coefficient**	**Std. Error**
Price/log(income)	-0.29313	0.82023	-0.33907	0.05713
Range	0.00396	0.00031	0.00669	0.00093
Acceleration	-0.07894	0.01378	-0.11652	0.02187
Top Speed	0.00422	0.00087	-	-
Pollution	-0.55782	0.10221	-0.75645	0.18203
Size	0.1276	0.03207	0.22116	0.0628
Luggage space	-	-	1.12805	0.41114
Operating cost	-0.10088	0.01064	-0.25231	0.03383
Station availability	0.27699	0.07455	0.70534	0.19206
Sports utility vehicle	0.86253	0.14617	0.92437	0.14968
Sports car	0.67947	0.15956	0.71357	0.16388
Station Wagon	-1.48132	0.06642	-1.51967	0.06782
Truck	-1.05403	0.05525	-1.11808	0.05592
Van	-0.80282-	0.05419	-0.81443	0.05619
Commute < 5 & EV[1]	-	-	0.42306	0.19038
College & EV	-	-	0.93633	0.25907
College & Methanol	-	-	0.39795	0.13779
CNG[2]	-	-	-0.08632	0.19047
EV	-	-	-1.35161	0.49765
Methanol	0.38902	0.05059	0.49892	0.17595
Random Effects				
Acceleration	0.20188	0.07144	-	-
Size	-	-	0.84084	0.25975
Operating cost	0.26487	0.03276	0.65276	0.10672
CNG	-	-	3.25604	0.61779
EV	-	-	2.95643	0.60767
Log likelihood	-7405.8		-7363.11	
BIC	14946.33		14920.45	

[1]Electric vehicle [2]Compressed natural gas

The variable known as Big Enough, previously included in the MAT model, was not included in Models 1a and 1b. A probable reason for this is that the effect of this variable could be explained by other variables. For example, the variables Size and Luggage Space can have information about whether a vehicle is big enough. Therefore, removing the variable Big Enough from the model does not have a large effect. A possible disadvantage for Models 1a and 1b can be the values of the random effects. These values are small compared to those in the MAT model. This can be inconvenient because they can be interpreted as nonsignificant random effects. In addition, Model 1a removes several variables that the analyst might consider important for the interpretation of the model.

The signs for the coefficients in Model 1a and 1b match the ones in MAT model, also, the magnitude of the coefficients is similar. The previous means that the overall effect of the variables on the output is similar for MAT model and Model 1a and 1b which leads to conclude that the models found by the proposed algorithm are meaningful and useful. For example, the variable Price has a negative sign which can be interpreted as: larger values for prices have a negative impact for the choice of a vehicle. In the other hand, the variable range has a positive sign with means that vehicles with larger values for range are preferred by decision makers. For the variables that are modeled as random parameters, the coefficients provide more insights about the preference of the decision makers. For example, for variable Electric Vehicle (EV), the coefficient -1.35 represents that, because of the negative sign, in average, people avoids this type of vehicles. However, the value of 2.9 of standard deviation represents that despite of the preference for Non-Electric Vehicles there is a big fraction of people who are willing to use electric vehicles. Probabilities above and below zero for the given mean and standard deviation following a normal distribution can be used to calculate the amount of people who like and dislike Electric Vehicles.

Computing these probabilities, it is possible to determine that 68% of the decision makers prefer Non-Electric Vehicles and the remaining 32% prefer Electric Vehicles.

As shown in Table 4, the MAT model has a BIC of 14,962.72 which is less than the BIC for Model 1. However, the likelihood ratio shows that the difference between these two models is significant. Therefore, compared to Model 1a, the MAT model fits the data better. On the other hand, a likelihood ratio test showed that Model 1b fit the data better than the MAT model. Even though the log likelihood of Model 2 was a little bit smaller, it was obtained using fewer parameters compared to MAT model. Hence, the difference in the log likelihood does not seem significant. The log likelihood ratio and the BIC provided evidence that the proposed algorithm could find a quality model in terms of goodness of fit.

Table 4. Summary of Quality Measures for Models

Model	BIC	Log-Likelihood
Dataset for alternative-fueled vehicles		
McFadden & Train (2000)	14962.7	-7358.9
Model 1a	14946.3	-7405.8
Model 1b	14920.4	-7362.9
Dataset for video streaming services		
Glasgow & Butler (2017)	8864.7	-4363.5
Model 2	8958.8	-4426.7

Experiment 3

Figure 4 illustrates the improvement in the BIC for Experiment 3. The initial BIC was 9826.08 and the final BIC was 8958.85. The behavior of the BIC through the iterations of the algorithm suggests that convergence was reached.

Table 5 shows Model 2, which is the output of the proposed algorithm for the third experiment. The random effects for variables Fast Content, More Content and, No Service follow

a normal distribution; and for variables Share NPII and PII, Price, and, Commercials follow a triangular distribution. The variable More TV/fewer movies initially included by Glasgow & Butler (2017) was not included by the proposed algorithm. A probable reason for this is that the inclusion of the variable More Content might be enough to explain the effect of the omitted variable.

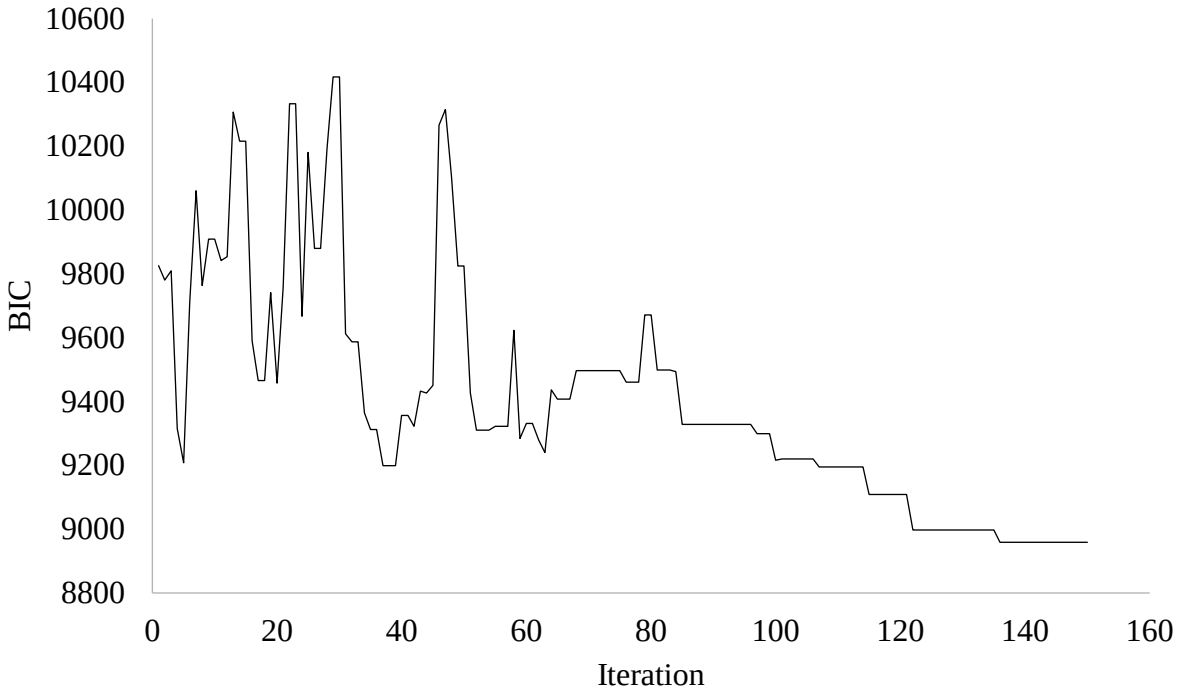

Figure 4. BIC vs iterations for Model 2.

Table 5. Algorithm output for Experiment 3.

Model 2		
Variable	**Coefficient**	**Std. Error**
Share NPII	-0.43209	0.053979
Share NPII and PII	-0.74832	0.069543
Price	-0.2342	0.013359
Commercials	-0.27574	0.047222
Fast Content	0.473953	0.048558
More Content	0.412229	0.049899
No Service	-3.36217	0.18228
Random Effects		
Share NPII and PII	2.06521	0.199235
Price	0.346337	0.015394
Commercials	1.42009	0.126587
Fast Content	0.66417	0.07209
More Content	0.74492	0.073734
No service	2.550361	0.15175
Log likelihood	-4426.76	
BIC	8958.854	

For Model 2, the signs of the coefficients are the same as the ones for the model originally proposed by Glasgow & Butler (2017), also the magnitudes of the coefficients are similar. The interpretation of these coefficients evidence that the effects of the variables is the expected considering preferences of people. For example, attributes such as Share Information, Price, and Commercials affect negatively the choice of a video streaming service; and attributes such as Fast Content and More Content affect positively the choice. These effects make sense in reality. The random effects for some of the coefficients allow a better understanding of the distribution of the preferences. For example, for the variable commercials the coefficient of -0.27 shows an average preference for services without commercials. However, the standard deviation value of 1.42 shows that this preference is dispersed and a significant share of the population is willing to pay for video streaming services with commercials. Using this mean and standard deviation it is possible to stablish that approximately 57% of the respondents to the survey prefer video streaming services without commercials and the remaining 43% are willing to accept commercials.

The improvement in the BIC and a likelihood ratio test evidence that the final model is a good quality model, however the goodness of fit is not as good as the one for the model originally proposed by Glasgow & Butler (2017) as shown in Table 4. The reason for this is that Glasgow & Butler (2017), using their knowledge in the data and the interpretation that they expected for the model, transformed the probability function to accommodate their analysis needs. The algorithm proposed on this study, does not apply transformations to neither to the data nor the probability function.

CHAPTER 6: CONCLUSIONS

The results suggest that the proposed algorithm can find an adequate specification for a mixed logit model in terms of goodness of fit. However, it is necessary to consider the judgement of the analyst in order to avoid suppression of variables or random effects important for the interpretation of the model. This can be handled by adding constraints to guarantee the inclusion of elements defined by the analyst. The main challenge when applying the proposed algorithm with a new dataset is to define the neighborhood criteria and cooling schedule for the simulated annealing algorithm. A single definition of these elements that can be applied to all problems does not exist. However, the existing literature provides references for this purpose. It is important to highlight that the proposed algorithm minimizes the intervention and required time of the analyst for the speciation of a mixed logit model. The algorithm only requires an initial configuration and even though it takes some hours to run, at the end of the process, the analyst obtains a model specification with substantial goodness of fit. This constitutes the proposed algorithm as a valuable tool to help analysts, with different levels of expertise in statistics, to specify mixed logit models.

The first experiment found a model specification with relatively small BIC. However, the likelihood ratio test was more favorable for the MAT model. In the second experiment, the proposed algorithm found a better model specification in terms of BIC and the log likelihood ratio test relative to the MAT model. This result was based on an ideal initial solution and illustrates the existence of better solutions which can potentially be obtained using an extensive search algorithm. Alternatively, an analyst could obtain Model 1b by first estimating the MAT model using their understanding of the problem and then applying the proposed algorithm to exploit the search space in the vicinity of such initial solution. This gives an opportunity for the analyst to pass valuable problem-specific knowledge to the algorithm. The fact that an algorithm can combine exploration

and exploitation could be more efficient when solving the problem formulation in this study. This is because the proposed optimization problem generally has a search space that is big; at the same time, small differences could substantially impact the objective function. A memetic algorithm is a metaheuristic which combines exploitation and exploration and is promising to solve the proposed problem regardless of the initial solution.

The proposed algorithm can be enhanced in future research to maximize the quality of the final model by including computations for overfitting, multicollinearity, and predictive performance. Other quality measures – such as prediction rate, Akaike information criteria, precision, and recall – can be used as objective functions. Also, McFadden & Train (2000) propose a test with artificial variables that helps to determine what variables can be modeled with random coefficients. This can be included in the proposed algorithm to reduce the search space by trying various density functions only for the coefficients specified by the artificial variables test. Also, the objective function could include a measure that penalizes random effects with low magnitude. Additionally, other metaheuristics, such as genetic algorithms or particle swarm optimization, that have been proven to be effective in optimization problems, can be applied to solve the proposed problem formulation.

Finally, transformations in the data and the probability function of mixed logit can be included as an additional optimization dimension for the algorithm. The previous can result in better model specifications. The authors who originally worked with the datasets used in this study, proposed good-quality specifications for mixed logit models by applying transformations to the data or to the structure of the probability function of mixed logit. For this purposed, they used their knowledge about the datasets and the context of the problem. In general, an approach that maximizes the inclusion of knowledge of the author about the problem and the data will represent

an improvement to the search ability of the proposed algorithm.

APPENDIX A: INSTRUCTIONS TO EXECUTE ALGORITHM

Requirements

- Operating system: Windows, Linux, Mac
- R version: 3.3.2
- Libraries: mlogit for R
- For better performance, a processor with speed superior to 3.2 Ghz is recommended.

Steps

1. Install R 3.3.2 (https://cran.r-project.org/bin/windows/base/old/3.3.2/R-3.3.2-win.exe)
2. Open R console
3. Install package mlogit for R using the following command:

   ```
   Install.Packages("mlogit")
   ```

4. Set working directory to the location of the folder of the experiment to be executed, using command setwd in the following way:

   ```
   setwd("c:\\Users\\Experiments\\Experiment1\\")
   ```

   Replace the path inside the quotes with path of the experiment folder in the local computer. Use \\ instead of \ for path separators in windows.
5. Open and execute the file mxlogit_search. R for the selected experiment.
6. During the process of execution, the console shows the progress through the iterations and a plot of BIC vs iterations is also shown. When the script stops, the output of the algorithm is stored in a file named 'mxlogit_out.txt' inside the experiment folder.

APPENDIX B: SOURCE CODE FOR ALGORITHM

Files Structure

The algorithm is organized in two main script files and one additional file with the parameters for a specific experiment. A description of the script files and their functionalities is provided below.

- **mxlogit_search.R** : This is the main script file. Global parameters, logging system and steps of simulated annealing algorithm are in this file. To run an experiment this is the file that must be executed. A regular user (not developer) should not modify this file.
- **mxlogit_search_fun.R**: Contains al the functions or methods used in the main file. Simulated annealing methods and some utility functions for logging are part of this file. A regular user (not developer) should not modify this file.
- **params.R**: Script file with all the parameters for a particular experiment. To use the algorithm with a new dataset, this is the file that the analyst must modify. In this file, the analyst must read the dataset and parse it to a R dataframe. The variables that the analyst want to be part of the analysis must be listed in the array ‘vars’. The variables that are alternative specific can be specified with the vector ‘asvars’. Same for individual specific variables ‘isvars’. The variables that need transformation for log normal distributions can be specified using the vector ‘lnvars’. The variables that the analyst does not want as random parameters can be specified using the array ‘fdvars’. All these arrays use the position in the array ‘vars’ as reference for the positions. Here 1 means enable and 0 disable. At the end of the this file the parameters for the simulated annealing are listed.

Source Code

mxlogit_search.R

```
library(mlogit)

#=====================================
#ENVIRONMENT
#=====================================
out_file = paste("mxlogit_out.txt",sep = "")
source("mxlogit_search_fun.R")
source("params.R")
cat("D  \tS  \tF \thits \tAIC     \tBIC     \t\tLL    \t\tsvars
\t\tfvars",file=out_file,sep="\n",append=TRUE)

#General parameters
rem_nonsig_coeff = TRUE
R = 100 #Number of random draws

#=====================================
#SIMULATED ANNEALING
#=====================================
start.time = Sys.time()
print(paste("Starting algorithm at: ",start.time))
all_M = list()
all_M_eval = list()
M = generate_initial_solution()
M = list(svars = svars, fvars = fvars)
M_eval = evaluate(M)
#------ Simulated annealing
Temp = Tini
iter = 1
repeat{
  #----- Generate Neighbor
  neighbors = lapply(1:NN,function(i) generate_neighbor(M))
#Generate NN neighbors
  evals = lapply(1:NN,function(i) evaluate(neighbors[[i]]))
#Evaluate NN neighbors
  Mc = neighbors[[which.min(evals)]]
  Mc_eval = evals[[which.min(evals)]]

  #---- Determine acceptance of neighbor
  if(Mc_eval < M_eval){ #Accept new neighbor as current solution
    M = Mc
    M_eval = Mc_eval
  }else{
```

```
    ap = acceptance_probability(M_eval,Mc_eval,Temp)
    if(runif(1, 0, 1) < ap){ #Check acceptance probability
      M = Mc
      M_eval = Mc_eval
    }
  }

  #---- Display/Store iteration findings
  print(paste("(",iter,")",M_eval))
  all_M[[iter]] = M
  all_M_eval[[iter]] = M_eval
  plot(unlist(all_M_eval),type = "l")

  #----- Update for next iteration
  Temp = Temp*cool_rate
  iter = iter + 1
  if(Temp < Tmin){break;}
}

print(paste("Finishing algorithm at: ",Sys.time()))
Sys.time() - start.time

#=====================================
#PRINT OUTPUT FILE
#=====================================
cat("Variables: ", vec2str(vars)
,file=out_file,sep="\n",append=TRUE)
cat("Alternative Specific Vars: ",
vec2str(asvars),file=out_file,sep="\n",append=TRUE)
cat("Vars with log transf.: ", vec2str(lnstvars)
,file=out_file,sep="\n",append=TRUE)
cat("\n","Evaluation / Models:
",file=out_file,sep="\n",append=TRUE)
cat(unlist(lapply(seq_along(all_M),function(i){paste(all_M_eval[
[i]],"\t", M2str(all_M[[i]]) )
})),file=out_file,sep="\n",append=TRUE)
plot(unlist(all_M_eval),type = "l")
```

mxlogit_search_fun.R

```
#=====================================
#FUNCTIONS
#=====================================
get_rand_density = function(current){
# Returns a random density different from the current one
  densities = c("t","n","ln")
  opts = densities[densities != current]
  pos = sample(1:length(opts), 1,replace=TRUE)
  return (opts[pos])
}


evaluate = function(M){
#Preproces and run mixed logit for specification M
  ev = 10000000 #Set high when minimizing
  error = TRUE
  #------- Transform data for lognormal cases
  TrainDataTmp = TrainData
  tvars = vars[M$svars == 1 & M$fvars == "ln" & lnstvars == 1]
#Variables to be transformed
  for(var in tvars){TrainDataTmp[var] = -TrainDataTmp[var]}
#Transform data

  #------- Mixed Logit execution
  fla = create_formula(M)
  print(paste("MxLogit: fla=
",paste(fla$formul[2],fla$formul[3],sep=' ~ '),";
rpars=(",paste(names(fla$rpars),"=",fla$rpars,collapse=","),");
svars=(",paste(M$svars,collapse=","),");
fvars=(",paste(M$fvars,collapse=","),");",sep=""))
  try({

  mxlogit = mlogit(fla$formul, TrainDataTmp, rpar = fla$rpars,
panel = is_panel, reflevel = reflev,  halton = NA, R = 20)
  rm(list=".Random.seed", envir=globalenv()) #Reset randoms
  deg_fre = length(mxlogit$coefficients)
  #compute_performance(mxlogit,deg_fre,"Original")
  if(rem_nonsig_coeff){
  #------- Remove non significant variables
    pvals = summary(mxlogit)$CoefTable[,4] #extract p-values
    non_sig = names(pvals[pvals > 0.09])   #non significant
variables
    mxlogit$coefficients[non_sig] = 0      #ignore non-
significant coefficients
```

```
    mxlogit$coefficients[match(paste("sd.",non_sig, sep =
""),names(mxlogit$coefficients))] = 0 #ignore nonsig
    deg_fre = length(mxlogit$coefficients) - length(non_sig)
    mxlogit <- update(mxlogit, start = coef(mxlogit), data =
TrainDataTmp, iterlim = 0, print.level = 0)
    rm(list=".Random.seed", envir=globalenv()) #Reset randoms
  }

  ev = compute_performance(mxlogit,deg_fre,"")

  error = FALSE
  })
  if(error){cat(paste("ERROR
with:",vec2str(M$svars),"\t",vec2str(M$fvars)),file=out_file,sep
="\n",append=TRUE);rm(list=".Random.seed", envir=globalenv()) }
  return (ev)
}


compute_performance = function(mxlogit,deg_fre,tag=""){
#Computes and logs predictive performance
  pred =
apply(mxlogit$probabilities,1,function(x){names(which.max(x))})
  pred[sapply(pred,is.null)] = "None"    #Mark null values as
None
  pred = unlist(pred)
  hits = sum(sapply(1:N,function(i){pred[i] == choices[i]}))

  rAIC = round(  2*deg_fre - 2*mxlogit$logLik  , digits = 3)
  rBIC = round(  log(length(choices))*deg_fre - 2*mxlogit$logLik
, digits = 3)

  evastr = paste(deg_fre ,"\t",sum(M$svars),"\t",sum(M$fvars !=
""),"\t",hits,"\t",rAIC,"\t",rBIC,"\t\t",round( mxlogit$logLik,
digits = 5),"\t\t",
vec2str(M$svars),"\t\t",vec2str(M$fvars),"\t",tag,sep = "")
  cat(evastr,file=out_file,sep="\n",append=TRUE)
  print(paste("hits=",hits,";   BIC=",rBIC ))
  return (rBIC)
}


create_formula = function(M){
# Creates the mixed logit formula for model specification M
  sel_asvars = M$svars==1 & asvars==1
  sel_isvars = M$svars==1 & isvars==1
```

```
  formul = formula(paste(
    paste(outcome, " ~ "),
    ifelse(sum(sel_asvars) > 0,paste(vars[sel_asvars],collapse =
" + "),"0"), #selas
    "|",
    ifelse(sum(sel_isvars) > 0,paste(vars[sel_isvars],collapse =
" + "),"0")  #selis
    ))

  rpars = setNames(M$fvars,vars)
  rpars = rpars[sel_asvars == 1] #Only for selected variables
  rpars = rpars[rpars != ""]
  return(list(formul = formul,rpars = rpars))
}


is_valid_neighbor = function(Mn){
#Check validity of a neighbor
  #At least 1 variable
  if(sum(Mn$svars) < 1) {return (FALSE)}
  #At least one alternative specific variable
  if(sum(Mn$svars==1 & asvars==1) < 1) {return (FALSE)}
  #At least one selected variable with density function
  if(sum(Mn$svars==1 & Mn$fvars!="") < 1) {return (FALSE)}
  return (TRUE)
}


generate_neighbor = function(M){
#Generates a neighbor solution
  repeat{ #until a valid neighbor is generated
    Mn = M

    #alter selected variables svars
    num_alterations = round(perc_alter_svars*length(vars))
    num_alterations = ifelse(num_alterations <
1,1,num_alterations) #at least 1 alteration
    positions = sample(1:length(vars),num_alterations,replace =
FALSE)
    old = Mn$svars[positions]
    Mn$svars[positions] = as.numeric(!old) #Update positions as
negation of old values

    #alter density functions fvars
    avail_pos = which(asvars == 1 & Mn$svars == 1 & fdvars != 1)
    num_alterations = round(perc_alter_fvars*length(avail_pos))
```

```
    num_alterations = ifelse(num_alterations <
1,1,num_alterations) #at least 1 alteration

    rand_pos =
sample(1:length(avail_pos),num_alterations,replace = FALSE)
    positions = avail_pos[rand_pos]
    Mn$fvars[positions] =
sapply(Mn$fvars[positions],function(x){ ifelse(x=="","n","")
})

    #Change distrbution for D positions
    avail_pos = which(Mn$fvars != "" & Mn$svars == 1 & fdvars !=
1)
    rand_pos =
sample(1:length(avail_pos),num_alterations,replace = FALSE)
    positions = avail_pos[rand_pos]
    #For each position get random density
    Mn$fvars[positions] =
sapply(Mn$fvars[positions],function(x){get_rand_density(x)})

    if(is_valid_neighbor(Mn)) {break}
  }
  return(Mn)
}

generate_initial_solution = function(){
#Generates random initial solution
  svars = rep(0,length(vars))
  fvars = rep("",length(vars))
  pos = sample(1:length(vars),length(vars)*0.9,replace = FALSE)
  svars[pos] = 1
  fvars[pos] =
sapply(fvars[pos],function(x){get_rand_density(x)})
  M = list(svars = svars, fvars = fvars)
  return (M)
}

acceptance_probability = function(M_eval, Mc_eval, Temp){
#Checks acceptance probability given difference in evaluations
  return ( exp(-(abs(M_eval-Mc_eval)/Temp*boltz )) )
}


M2str = function(M){
#Returns a string with elements of model M
  return (paste("S =",paste(M$svars,collapse=","),"    F
=",paste(M$fvars,collapse=",") ))
```

```
}

vec2str = function(vec){
#Returns a string with elements of array vec
  return (paste(vec,collapse=","))
}
```

params.R

```
#=====================================
#DATASET PARAMETERS
#=====================================
Car = read.csv("Car.csv")

CarLong <- mlogit.data(Car, shape = "wide", varying = 2:139,
choice = "choice",sep = "")

Data = Car            #Data in wide format
TrainData = CarLong  #Data in long format
choices = TrainData[TrainData$choice,]$alt #Vector of choices
N = length(choices)
outcome = "choice"
reflev = "1"
vars =
c("price","range","acc","speed","pollution","size","be","space",
"cost","station","suv","sport","wagon","truck","van","ev","comlf
ive","colev","cng","methanol","colnmethan")#variable names

asvars = c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
#Alternative specific variables
isvars = as.numeric(!asvars)
#Individual specific variables
fvars =
c("n","","","","","","","","","","","","","","","","","","","
") #Distribution for alternative specific variables
svars = c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
#Selected variables
fdvars = c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
#Variables with fixed distrubution function
lnstvars = c(1,0,1,0,1,0,0,0,1,0,0,0,1,1,1,1,0,0,0,0,0)
#Variables that need sign to be transformed when log normal
is_panel = FALSE
if(! (length(vars) == length(fvars) && length(vars) ==
length(asvars) && length(vars) == length(svars)) ){
  stop("Size of vectors associated with variables must match")
}

#=====================================
#SIMULATED ANNEALING PARAMETERS
#=====================================
perc_alter_fvars = 0.18   #Alteration percentage for densities
perc_alter_svars = 0.18   #Alteration percenta for selected
variables
NN = 3                    #Number of neighbors
```

```
Tini = 1                    #Initial temperature
Tmin = 0.0022               #Final temperature
cool_rate = 0.96            #Cooling rate
boltz = 0.0004              #Boltzman constant
```

REFERENCES

Balcombe, K., Chalak, A., & Fraser, I. (2009). Model selection for the mixed logit with Bayesian estimation. *Journal of Environmental Economics and Management*, *57*(2), 226–237. https://doi.org/10.1016/j.jeem.2008.06.001

Ben-Akiva, M., & Lerman, S. (1985). *Discrete choice analysis: theory and application to travel demand*.

Brownstone, D. (2000). Discrete Choice Modeling for Transportation. *Travel Behavior Research: The Leading Edge*, (July), 97–124.

Brownstone, D., & Train, K. (1998). Forecasting new product penetration with flexible substitution patterns. *Journal of Econometrics*, *89*(1–2), 109–129. https://doi.org/10.1016/S0304-4076(98)00057-8

Brusco, M. J. (2014). A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. *Computational Statistics and Data Analysis*, *77*, 38–53. https://doi.org/10.1016/j.csda.2014.03.001

Brusco, M. J., & Steinley, D. (2011). Exact and approximate algorithms for variable selection in linear discriminant analysis. *Computational Statistics & Data Analysis*, *55*(1), 123–131. https://doi.org/10.1016/j.csda.2010.05.027

Changpetch, P., & Lin, D. K. J. (2013). Selection of multinomial logit models via association rules analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(1), 68–77. https://doi.org/10.1002/wics.1242

Croissant, Y. (2012). Estimation of multinomial logit models in R : The mlogit Packages An introductory example. Retrieved from https://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf

Daniel McFadden and Kenneth Train. (2000). Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, *15*(5), 447–470.

Efromyson, M. (1960). Multiple Regression Analysis. Mathematical Methods for Digital Computers. *John Wiley and Sons, Inc. NewYork*, 65–79.

Fouskakis, D., & Draper, D. (2008). Comparing Stochastic Optimization Methods for Variable Selection in Binary Outcome Prediction, With Application to Health Policy. *Journal of the American Statistical Association*, *103*(484), 1367–1381. https://doi.org/10.1198/016214508000001048

Glasgow, G., & Butler, S. (2017). The value of non-personally identifiable information to consumers of online services : evidence from a discrete choice experiment services : evidence from a discrete choice experiment. *Applied Economics Letters*, *24*(6), 392–395. https://doi.org/10.1080/13504851.2016.1197357

Hajek, B. (1988). Cooling Schedules for Optimal Annealing. *Mathematics of Operations Research*. INFORMS. Retrieved from http://www.jstor.org/stable/3689827

Hasan Örkcü, H. (2013). Subset selection in multiple linear regression models: A hybrid of genetic and simulated annealing algorithms. *Applied Mathematics and Computation*, *219*(23), 11018–11028. https://doi.org/10.1016/j.amc.2013.05.016

Hensher, D. A., & Greene, W. H. (2003). The mixed logit model: The state of practice. *Transportation*, *30*(2), 133–176. https://doi.org/10.1023/A:1022558715350

Hensher, D. A., & Ton, T. T. (2000). A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. *Transportation Research Part E: Logistics and Transportation Review*. https://doi.org/10.1016/S1366-5545(99)00030-7

Kadane, J. B., & Lazar, N. A. (2004). Methods and Criteria for Model Selection. *Journal of the*

American Statistical Association, *99*(465), 279–290. https://doi.org/10.1198/016214504000000269

Karlaftis, M. G., & Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research : Differences , similarities and some insights. *Transportation Research Part C*, *19*(3), 387–399. https://doi.org/10.1016/j.trc.2010.10.004

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science, New Series*, *220*(4598), 671–680.

Lin, S.-W., Lee, Z.-J., Chen, S.-C., & Tseng, T.-Y. (2007). Parameter determination of support vector machine and feature selection using simulated annealing approach. https://doi.org/10.1016/j.asoc.2007.10.012

Luo, G. (2015). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *5*(1), 18. https://doi.org/10.1007/s13721-016-0125-6

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010

Meiri, R., & Zahavi, J. (2004). Using simulated annealing to optimize the feature selection problem in marketing applications. https://doi.org/10.1016/j.ejor.2004.09.010

Milton, J. C., Shankar, V. N., & Mannering, F. L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis & Prevention*, *40*(1), 260–266. https://doi.org/10.1016/j.aap.2007.06.006

Nourani, Y., & Andresen, B. (1998). A comparison of simulated annealing cooling strategies. *J. Phys. A: Math. Gen*, *31*(98), 8373–8385. Retrieved from

http://www.fys.ku.dk/~andresen/BAhome/ownpapers/permanents/annealSched.pdf

Omrani, H. (2015). Predicting Travel Mode of Individuals by Machine Learning. *Transportation Research Procedia*, *10*, 840–849. https://doi.org/10.1016/j.trpro.2015.09.037

Pacheco, J., Casado, S., & Núñez, L. (2009). A variable selection method based on Tabu search for logistic regression models. *European Journal of Operational Research*, *199*(2), 506–511. https://doi.org/10.1016/j.ejor.2008.10.007

Pacheco, J., Casado, S., & Porras, S. (2013). Exact methods for variable selection in principal component analysis: Guide functions and pre-selection. *Computational Statistics & Data Analysis*, *57*(1), 95–111. https://doi.org/10.1016/j.csda.2012.06.014

Paz, A., Emaasit, D., & de la Fuente, H. (2016). Stochastic dynamic user equilibrium using a mixed logit modeling framework. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1206–1211). IEEE. https://doi.org/10.1109/ITSC.2016.7795710

Paz, A., Molano, V., Martinez, E., Gaviria, C., & Arteaga, C. (2015). Calibration of traffic flow models using a memetic algorithm. *Transportation Research Part C: Emerging Technologies*, *55*, 432–443. https://doi.org/10.1016/j.trc.2015.03.001

Paz, A., & Peeta, S. (2009). Paradigms to Deploy a Behavior-Consistent Approach for Information-Based Real-Time Traffic Routing. *Networks and Spatial Economics*, *9*(2), 217–241. https://doi.org/10.1007/s11067-008-9077-4

Pinjari, A. R., Pendyala, R. M., Bhat, C. R., & Waddell, P. A. (2011). Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, *38*(6), 933–958. https://doi.org/10.1007/s11116-011-9360-y

Sato, T., Takano, Y., Miyashiro, R., & Yoshise, A. (2016). Feature subset selection for logistic regression via mixed integer optimization. *Computational Optimization and Applications*, *64*(3), 865–880. https://doi.org/10.1007/s10589-016-9832-2

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sutter, J. M., & Kalivas, J. H. (1993). Comparison of Forward Selection, Backward Elimination, and Generalized Simulated Annealing for Variable Selection. *Microchemical Journal*, *47*(1–2), 60–66. https://doi.org/10.1006/mchj.1993.1012

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511753930

Tutz, G., Pößnecker, W., & Uhlmann, L. (2015). Variable selection in general multinomial logit models. *Computational Statistics & Data Analysis*, *82*, 207–222. https://doi.org/10.1016/j.csda.2014.09.009

Vicari, D., & Alfó, M. (2014). Model based clustering of customer choice data. *Computational Statistics and Data Analysis*. https://doi.org/10.1016/j.csda.2013.09.014

Vinterbo, S., & Ohno-Machado, L. (1999). A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction. *Proceedings. AMIA Symposium*, 984–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10566508

Wegener, M. (2004). Overview of Land Use Transport Models (pp. 127–146). https://doi.org/10.1108/9781615832538-009

Zahid, F. M., & Tutz, G. (2013). Multinomial logit models with implicit variable selection. *Advances in Data Analysis and Classification*, *7*(4), 393–416. https://doi.org/10.1007/s11634-013-0136-4

CURRICULUM VITAE

Graduate College

University of Nevada, Las Vegas

Cristian David Arteaga Sanchez

Degrees:

Bachelor in Computer Science, 2015

Universidad del Cauca, Colombia

Publications:

Paz, A., Molano, V., Martinez, E., Gaviria, C., & Arteaga, C. (2015). Calibration of traffic flow models using a memetic algorithm. *Transportation Research Part C: Emerging Technologies, 55*, 432-443.

Cobos, C., Daza, C., Martínez, C., Mendoza, M., Gaviria, C., Arteaga, C., & Paz, A. (2016). Calibration of CORSIM Vehicular Traffic Flow Models using a Memetic Algorithm with Solis and Wets Local Search Chaining. *Advances in Artificial Intelligence: Lecture Notes in Computer Science.* 10022, 365-375.

Cobos, C., Erazo, C., Luna, J., Mendoza, M., Gaviria, C., Arteaga, C., & Paz, A. (2016). Multi-Objective Memetic Algorithm based on NSGA-II and Simulated Annealing for Calibrating CORSIM Micro-Simulation Models of Vehicular Traffic Flow. *Advances in Artificial Intelligence: Lecture Notes in Computer Science*. 9868, 468-476.

Paz A., Arteaga, C. Specification of Mixed Logit Models Using an Optimization Approach. In *Transportation Research Board 97nd Annual Meeting. Under Review*

Veeramisti, N., Paz, A., Khadka, M., & Arteaga, C. Estimation of Safety Performance Functions Using Clusterwise Regression. *Accident Analysis and Prevention. Under Review*

Khadka, M., Paz, A., Arteaga, C., & Hale, D. Simultaneous Generation of Optimum Pavement Clusters and Associated Performance Models. *Journal of Infrastructure Systems. Under Review*

Shrestha, K., Paz, A., Arteaga, C. & Gaviria, M. Calibration of microscopic traffic flow simulation models enabling selection of links and parameters simultaneously. In *Transportation Research Board 97nd Annual Meeting. Under Review*

Thesis Title:

Specification of Mixed Logit Models Using an Optimization Approach

Thesis Examination Committee:

Chairperson, Alexander Paz, Ph.D.

Committee Member, Dave James, Ph.D.

Committee Member, Mohamed Kaseko, Ph.D.

Committee Member, Brendan Morris, Ph.D.

Graduate College Representative, Justin Zhan, Ph.D.