|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | | Term Project Report Part 1 | | | | |  | |
|  |  | | | | | | |  |
|  | | | |  |  | | | |
|  | | | | Tatum Edge **Akhilesh Shastri**  **Nicholas Runde** |  | | | |
|  | | | | Date: 11/10/20—Transportation Data Analytics—Dr. Xilei Zhao |  | | | |
|  | | |  | | |  | | |

Contents

[Introduction: 3](#_Toc55491138)

[Literature Review: 4](#_Toc55491139)

[Data Preparation and Preprocessing: 5](#_Toc55491140)

[Future: 7](#_Toc55491141)

### Introduction:

Being familiar with Transportation modeling, we will jump right into the type of modeling framework to be used. The sociological term *Substantive Rationality* indicates that, as human beings, we know what our objectives are and we can envisage all alternative ways of achieving them and, with some luck, quantify the costs and benefits associated to each approach. It can be applied to important decisions like choosing a place to live and even less important ones like choosing a place to eat. This is the rational or normative decision-making approach implicit in *Economics, Sociology, and Planning*. The same “rational decision-making approach” can be extended to choosing the mode of Transportation by the User (technically referred to as Mode Choices). To predict the mode choices of users, a modelling framework is used referred to as Modal Split or Route Choice Models the Logit Model is one such a type in this framework. The *Logit Model* uses an economical term named *Utility* which, in the theory of consumer behavior, is termed as ***“Consumer Preferences.”*** In short,the *Utility* defines a number and is seen as a way to describe user preferences. As far as choice behavior of the user is concerned what matters is whether *Utility* of one choice is higher than its alternative. (Garber. 2009)

Trips can be completed using different modes of travel. The proportion of trips undertaken by each of the different modes is termed modal split. The simplest form of modal split is between public transport and the private car.While modal split can be carried out at any stage in the transportation planning process, it is assumed here to occur between the trip distribution and assignment phases. The trip distribution phase permits the estimation of journey times/costs for both the public and private transport options. The modal split is then decided on the basis of these relative times/costs. The Logit Model considers the relative utilities of each mode as a summation of each modal attribute. The modal attributes are generally taken as IVTT (in vehicle travel time), OVTT (out of vehicle travel time), and COST (out of pocket cost for users). In order to simplify the computation of modal split, journey time is taken as the quantitative measure of the cost criterion. The decision by a commuter regarding choice of mode can be assumed to have its basis in the micro-economic concept of utility maximization. This model presupposes that a trip maker selects one particular mode over all others on the basis that it provides the most utility in the economic sense. One must therefore be in a position to develop an expression for the utility provided by any one of a number of mode options. The function used to estimate the total utility provided by a mode option usually takes the following form:(Rogers 2003)

Where,

U­i = utility function for mode i

x = IVTT (in vehicle travel time)

y = OVTT (out of vehicle travel time)

z = COST (out of pocket cost for users)

α, β, γ = Arbitrary parameters

Once the utilities of each mode are known the probability that a trip maker will select one mode option, m, is equal to the probability that this option’s utility is greater than the utility of all other options. This can mathematically formulated as:

Where,

Pm = Probability that mode m is chosen

Umi = Utility of ith mode among n different modes available.

The above model presupposes that a trip maker selects one particular mode over all others on the basis that it provides the most utility in the economic sense.(Lester 2009)

### Literature Review:

In a modernizing world, the demand for transportation is ever-expanding. Transportation engineers and planners have a vital job in maintaining the developing transportation systems. When evaluating the development and maintenance of a transportation network, transportation engineers and planners must effectively understand the demand within the network. Modern approaches to understand user demand, such as travel surveys, can help planners and engineers make adequate decisions for network optimization. To develop effective data regarding user demand in a network, a utility function is an essential evaluation. “Utility theory bases its beliefs upon individuals’ preferences” therefore, when evaluating a transportation network, an engineer or planner can understand the preferred mode choice for all users in the network (Saylor).

While the determination of mode-choice use is an essential requirement of transportation planning, it is not yet adequately handled in the process. If the factors perceived by a traveler are seen to be equal between modes (i.e. perceived "costs" are equal), then it is equally likely that either mode will be chosen. However, irrespective of costs, some travelers must use a particular mode which considers them"captive" to that mode (this goes beyond the scope of research covered in this paper). Research has indicated that the attitude of individuals to public transport can be measured and ranked in order of preference. The most valued travel attributes in order are arrival on time, getting a seat, no transfers, regularity of service, weather protection while waiting and shorter wait stop.(Hobbs. 1979)

Many models have been developed, usually based on empirical studies, leading to diversion curves, multiple regression and discriminant analysis using probability or least-squares techniques to fit the observed data. The London Traffic Survey (Coburn 1960) extended the 108 household categories by adding accessibility groups (low, medium, and high) for bus and train levels of service indices for a zone. The procedure, adopted by TRC in the Washington Study(Brand et al 1973), compiled 80 categories of trips and trip makers (4 categories of car ownership, 5 of income and 4 for service ratios) separating work and non-work journeys. Diversion curves were constructed to assign the public transport users as a function of travel time, cost, and service ratios, stratified by the traveler’s economic status. The relative convenience of modes was assessed by the service ratio, including excess time; the ratio of excess time spent on a public transport journey to that of the car was taken as the relative convenience factor in the service ratios (however, this research will not use service ratios for the inference). Public transport excess time included walking, waiting and transfer time while that for the car was the time spent walking to and from parking areas and parking time. Discriminant multiple regression methods have formed the main basis of disaggregate modelling by Warner, the logit model, Quarmby, and the LGORU model (that will not be used in this research). When disaggregate data is used the dependent variable (e.g. mode chosen) takes a value of 0 or 1 (not chosen/chosen), and the model coefficients are usually estimated by maximum likelihood methods. The values 0 or 1 are the inputs of a factored dummy variables, for example variables like “Passenger owns a car” (yes or no / 1 or 0) etc. Data used in transportation models can be measured from observation (or estimated) or perceived values can be used as reported by individuals. Differences between measured and reported data represent the difference between reality and as the individual perceives reality, the latter being influenced by both psychological and physiological attributes at a point in time. (Hobbs. 1979).

Daly (1997) discussed various models and concluded that nested logit models were most adequate. Hence to effectively predict mode choice by user of disintegrated model, this research will utilize a multi-modal nested logit model.

### Data Preparation and Preprocessing:

To effectively build a Utility Function that depicts the demand in a transportation network, a Household Travel survey is a vital tool. Household Travel Surveys are administered by either phone or mail to develop an understanding for the daily travel in a transportation network. New York City has a diverse transportation network that incorporates various modes of transportation. To maintain such a broad network, the city of New York must understand the demand within the system. Annually, the city of New York administers a Household Travel Survey to analyze trends within the transportation network. For the model development in this project, the data acquired by New York DOT for the Household Travel Survey 2018 will be utilized. Because the data is utilized by the local government for transportation analytics, the dataset is relatively clean. However, portions of the data are not beneficial for developing a predictionary model for mode choice. To ensure that only relevant data is considered, data pre-processing and processing must take place.

For the data, despite it being relatively clean, pre-processing must still occur to ensure that the garbage-in, garbage-out dilemma is not reached. The data must not contain any errors that could potentially skew the overall model. Therefore, pre-processing requires the removal of all NULL or NA values in the data. Due to the style of the questionnaire, some outputs automatically produce unwanted responses. On certain questions, like “"Which of the following car sharing services, if any, are you a member of?", the options to ‘Refuse’ and ‘Do not Know’ are considered. However, to perform analysis on the data, it is essential we only include questions that provide an answer from the participant. Therefore, these questions, along with the individuals (if any) that answer in such a way, will be best if removed from the data set.

In addition to the removal of irrelevant variables, it is important to consider what questions will be asked in the model. Categorical variables such as the questions regarding a ‘Disability’ can effectively be moved to one group. To accomplish this, ‘Disability’ is converted into a binary-variables, which registers as ‘Yes’ for participants that have a disability and ‘No’ for participants that do not have a disability. The development of binary variables should also be practiced for the questions regarding Car Sharing and Welfare. This provides a better understanding of how the presence of one of these three variables will, or will not, influence the model.

### Future:

Now that the variables are set and the obviously irrelevant variables are thrown out, the model development can begin. For model development, the utility function requires the development of a linear model for each mode choice. To develop a linear model for each mode choice, it is important to run analysis on each variable to see how they are related. Variables with high correlation must be monitored when developing the model. Variables with high correlation could potentially skew the model; therefore, one variable may need to be thrown out.

After analyzing the variables and their inter-relationships, tests, such as Best Subset Analysis will be useful in determining which independent variables act as the best predictors in a mode choice study. Once the ideal linear model is produced for each mode choice, a utility function, which determines the probability of choosing each mode, will be designed. From the utility model, the model should further be utilized to predict the travel behavior of a random commuter for any given attributes.

*References:*

* BRAND, D. and MANHEIM, M. (eds.) (1973) Urban Travel Demand Forecasting, TRB Special Report No. 143. Transportation Research Board, Washington DC.
* COBURN, T. M., BEESLEY, M. E. and REYNOLDS, D. J. (1960) The London-Birmingham Motorway: Traffic and Economics, Road Research Technical Paper No. 46. HMSO, London.
* Nicholas J. Garber and Lester A. Hoel (2009) Traffic and Highway Engineering, CENGAGE Learning.
* Martin Rogers (2003) Highway Engineering, Blackwell Science.
* F.D. Hobbs (1979) Traffic Planning and Engineering, Pergamon Press.
* Juan de Dios Ortuzar, Luis G. Willumsen (2011) Modelling Transport, WILEY
* Hal R. Varian (1987) Intermediate Micro Economics a modern approach.