Data-driven Modelling and Analysis of Household Travel Mode Choice

Nagesh Shukla, Jun Ma, Rohan Wickramasuriya, Nam Huynh

SMART Infrastructure Facility, University of Wollongong, NSW 2522, Australia
Email: nshukla@uow.edu.au

Abstract: One of the important problems studied in the area of travel behaviour analysis is travel mode choice which is one of the four crucial steps in transportation demand estimation for urban planning. State of the art models in travel demand modelling can be classified as trip based; tour based; and activity based. In trip based approach, each individual trips is modelled as independent and isolated trips i.e. no connections between different trips. In tour based approach, trips that start and end from the same location (home, work, etc) and trips within a tour are dependent on each other. In past two decades, researchers have focussed on activity based modelling, where travel demand is derived from the activities that individuals need/wish to perform. In this approach, spatial, temporal, transportation and interpersonal interdependencies (in a household) constrain activity/travel behaviour.

This paper extends tour-based mode choice model, which mainly includes individual trip level interactions, to include linked travel modes of consecutive trips of an individual. Travel modes of consecutive trip made by an individual in a household have strong dependency or co-relation because individuals try to maintain their travel modes or use a few combinations of modes for current and subsequent trips. Traditionally, tour based mode choice models involved nested logit models derived from expert knowledge. There are limitations associated with this approach. Logit models assumes i) specific model structure (linear utility model) in advance; and, ii) it holds across an entire historical observations. These assumptions about the predefined model may be representative of reality, however these rules or heuristics for tour based mode choice should ideally be derived from the survey data rather than based on expert knowledge/judgment. Therefore, in this paper, we propose a novel data-driven methodology to address the issues identified in tour based mode choice. The proposed methodology is tested using the Household Travel Survey (HTS) data of Sydney metropolitan area and its performances are compared with the state-of-the-art approaches in this area.

Table 1. Classification of state of the art approaches in mode choice

Data Type / Trip Type	Discrete Choice Models		Machine Learning	
	Crisp Data	**Crisp & Fuzzy Data**	**Crisp Data**	**Crisp & Fuzzy Data**
Independent Trips	Gaudry, (1980); McFadden (1973); Daly & Zachary (1979); Hensher & Ton (2000)	Dell'Orco *et al.* (2007)	Xie et al. 2003; Reggiani & Tritapepe 1998; Cantarella et al., 2003; Shmueli et al. 1996; Edara 2003; Hensher and Ton, 2000	Yaldi, G. (2005)
Linked Individual Trips (tour-based)	Miller et al. (2005)	-	Biagioni et al., (2008)	This Study
Linked Household Trips	Miller et al. (2005)	-	Future Work	Future Work

Keywords: *Travel mode choice, data mining, travel mode choice, fuzzy sets*

1. INTRODUCTION

One of the fundamental processes that shape urban landscapes is people's travel behaviour. Hence, a thorough understanding of travel behaviour is crucial for effective transportation and land use planning in urban environments. Travel mode choice is an important aspect of travel behaviour, and also one of the four steps in transportation demand estimation for urban planning. It refers to the procedure of assigning available travel modes (e.g. car, walk, bus, and train) to each individual's trips in a household based on personal, activity and environmental attributes.

Travel mode choice has received a significant research attention. From a modelling perspective, travel mode choice has been primarily studied using discrete choice models (reference). Such models include probit models (Gaudry 1980), multinomial logit (MNL) models (McFadden 1973) and nested logit models (Daly & Zachary 1979). However, discrete choice models have received stringent criticisms due to their inherent limitations such as i) specific model structure needs to be specified in advance, which ignores partial relationships between explanatory variables and travel modes for subgroups in a population; ii) inability to model complex non-linear systems, which represent complex relationships involved in human decision making; and iii) they check only for conditions that hold across an entire population of observations in the training dataset and patterns cannot be extracted from a subgroup of observations (Xie, et al., 2003).

Meanwhile, machine learning has emerged as a superior means in travel mode choice research by which travel mode choice can be better predicted while alleviating aforementioned shortcomings (Xie et al. 2003; Reggiani & Tritapepe 1998; Cantarella et al., 2003; Shmueli et al. 1996). For example, Xie, et al., (2003) report that Artificial Neural Networks (ANN) achieved better results compared to MNL based on a comparative study conducted using work-related travel data. Similarly, Isaradatta (2006) has compared nested logit model and ANN model for long distance travel mode choice selection, and illustrated the better performance of ANN over other models. Furthermore, there are other studies that compare and contrast the performance of machine learning techniques with other traditional models, and propose to use machine learning techniques such as ANN and DT for travel mode choice prediction (Edara 2003; Cantarella and de Luca, 2003; Reggiani and Tritapepe, 1998; Nijkamp, et al., 1998; Shmueli, et al., 1996; Hensher and Ton, 2000).

In addition to the differences in methods used, research into travel mode choice exhibits variations in terms of the predicted trip type, i.e., independent trips verses tour-based (linked) trips and data type used. It is important to understand these variations prior to establishing innovative aspects of this study. We identify three types of trips, (a) independent trips, (b) linked trips of an individual, and (c) linked trips of individuals within a household. To predict these trip types, researchers have used crisp or fuzzy data or a mix of crisp and fuzzy data. In this study, we use machine learning techniques to predict both independent trips and linked trips of an individual using a mix of crisp and fuzzy data. We are unaware of any other study where a mix of crisp and fuzzy data is used to for predicting the modes for tour-based linked trips. This methodological advance is rightly justified in results we achieved as explained in a later section. Table 1 serves three purposes: it summarizes existing research in travel mode choice, puts this study in perspective and identifies future research directions. The overall objective of this study is to achieve higher accuracy in travel mode choice predictions using machine learning algorithms.

The remainder of the paper is organised as follows.

2. PROPOSED MODELLING METHODOLOGY

This section details proposed modelling methodology in this paper for the travel mode choices of an individual in a household based on a travel survey. As mentioned in Section 1, travel mode choice problem has been studied largely using discrete choice models such as probit model, multinomial logit (MNL) model and nested logit models. Major limitations in these studies are i) predefined utility model with all the explanatory variables included ignoring partial relationships; ii) inability to model non-linear relationships; and, iii) models cannot be extracted from a subset of observations. This led us to explore methods in the area of machine learning, artificial neural networks (ANN) and decision trees (DT), to overcome the aforementioned limitations. These methods have predominantly been used for problems related to classification based on historical data or evidence. Following is the brief description of these methods:

Artificial neural networks (ANN): McCulloch and Pitts (1943) developed the concept of artificial neurons to study cognitive processes. Following this, past two decades saw major developments in the area of ANN application for variety of classification and pattern recognition problems in machine learning. In general, an ANN consists of set of interconnected processing nodes called neurons which is used to estimate mapping

between explanatory variables and the responses. Each processing nodes combines its inputs into a single output value based on activation function. Activation function first combines all the incoming inputs and then uses transfer functions to use combined inputs to produce single output value. Commonly used neural network for pattern recognition and classification consists of muti-layer feedforward (MLF) neural networks. It has been recognized that using multiple layers improves the ability of ANN to model complex linear as well non-linear relationships. Hence, we employ MLF ANN for this study.

Decision trees (DT): Decision trees are a rule based model which maps observations (explanatory variables) to their observed response. Due to its simple and intuitive nature, DT have used for classification in variety of areas in literature. DT resembles human reasoning and produces a white box model for decision making. In DT, leaves represent a class label and each node represents a condition or rule on explanatory variables. DT presents several advantages such as robustness to noise, low computational cost for model generation and ability to deal with redundant variables. DT algorithms such as C4.5 and Classification and Regression Technique (CART) have been identified as top 10 data mining algorithms in terms of its wider applicability (Wu et al. 2008).

Following section mainly presents the data processing and fuzzy analysis of dataset for modelling and prediction. These steps are used for ANN and DT for learning and prediction.

2.1. Data processing to link consecutive trips of an individual

Travel mode choice can be understood as the travel mode to which traveller pre-commit given a particular purpose (shopping, work, school) and other travel details (departure times, arrival times, origin, destination, etc). Majority of the traditional literature focuses on individual trip, i.e., the travel between a pair of origin and destination. In this type of modelling, each trip is considered as an independent event, for which an individual has to make independent decision about travel mode. However, due to complexity of patterns of trip of an individual, assumption about each trip to be independent does not hold well. Furthermore, Cirillo and Axhausen (2002) suggested that individuals maintain their mode during a tour (a sequence of trips starting and ending at the same place, i.e., home → work → shopping → home), especially if they use an individual vehicle (car, motorcycle or bicycle). Following the assumption that there is strong dependency between travel modes adopted in consecutive modes, this subsection considers consecutive trip modes (in a tour) for modelling and prediction.

Travel surveys are increasingly used in most of the metropolitan cities to understand the people's travel behaviour and demand for transport planning. These travel surveys record socio-economic characteristics, demographic characteristics, household attributes, travel details/diary, purpose, departing and arriving times, and travel modes, among others. These records are used by planner to design or change existing transport plans. This paper will utilize these travel surveys to model mode choices of an individual given other attributes.

Let $(\boldsymbol{X}, \boldsymbol{Y})$ be a survey dataset of trips made by L travelers, where $(\boldsymbol{x}^{lm}, \boldsymbol{y}^{lm})$ represents the m-th trip made by the traveller l, $m \in \{1, 2, \dots, M_l\}$, $l \in \{1, 2, \dots, L\}$. Without loss of generality, suppose $\boldsymbol{x}^{lm} = (x_1^{lm}, x_2^{lm}, \cdots, x_n^{lm})$ and $\boldsymbol{y}^{lm} = (y_1^{lm}, y_2^{lm}, \cdots, y_o^{lm})$ where each x_i $(i = 1, \dots, n)$ is called an explanatory attribute and each y_k $(k = 1, \dots, o)$ is a Boolean decision variable which indicates a possible travel mode.

In order to describe consecutive trips, we introduce an additional set of explanatory variables, which consider the mode choice adopted by the individual in previous trip, apart from $\boldsymbol{x}^{lm}$ to model the travel mode choices. The additional variable set is represented as $\boldsymbol{x}'_{lm}$ and it contains the mode choice of previous trip on a particular tour, i.e., $\boldsymbol{y}^{l(m-1)}$. Formally,

$$\boldsymbol{x}'_{lm} = \begin{cases} \mathbf{0} & m = 1 \\ \boldsymbol{y}^{l(m-1)} & m \in \{2, \dots, M_l\} \end{cases} \quad (1)$$

Since, the first trip (i.e., $m = 1$) does not have information about the previous trip mode choice ($\boldsymbol{y}^{l0}$), we use a dummy vector $\mathbf{0} = (0, \dots, 0)$ to represent that. In other words we treat first trip on an individual independent and the rest of the trips to be dependent on previous trip in a tour.

The modified travel survey dataset includes $\boldsymbol{x}^{lm}$ and $\boldsymbol{x}'_{lm}$ as an explanatory set of variables to model the responses in $\boldsymbol{y}^{lm}$ for all $l \in \{1, 2, \cdots, L\}, m \in \{1, 2, \cdots, M_l\}$. Now, this dataset is used with ANN and DT for modeling the travel mode choices of an individual in his tour.

Next subsection discusses the use of fuzzy data for modelling mode choices.

2.2. Fuzzy data for mode choice

Fuzzy set was introduced by Zadeh (1965) as a tool for processing uncertainty in real application systems which involve human perceptions of vague concepts such as "young person" and "big heap". Since then, fuzzy set has been successfully used in engineering, control systems, and decision making (Klir and Yuan 1995). Recently, it has been used in travel demand modelling (Yaldi, Taylor et al. 2010). Considering that a travel mode choice is a decision making on the basis of a set of uncertain factors including travel cost, travel time, purpose, as well as individual demographic characteristics, we argue that using fuzzy sets can better describe a person's choice of a specific travel mode.

Travel mode choice is affected by many uncertain factors. A typical factor is the travelling period. In Sydney metropolitan area, a traveller who drives to Central Business District (CBD) during the morning peak hours is very likely to experience traffic congestions and delays. However, if the traveller makes the same trip by train during the same period, the traffic congestion has minimal impact on this trip. A traveller's demographic characteristics may also affect his or her choice of a specific travel mode. A frequent traveller prefers driving a car to taking a public transport because of the flexibility the former offers for subsequent trips. In these examples, "morning peak hours" and "frequent traveller" are uncertain concepts whose meanings are easily understood but are hardly defined in an accurate way. Hence, using fuzzy set is an alternative to describe these uncertain concepts and related factors of them.

Fuzzy set can provide better description of and insight into a specific travel mode choice. Generally, a travel mode choice can be described as an "IF-THEN" expression such as: IF the depart time is 06:30 and the travel distance is 20.5km, THEN the travel mode is car-driving. Although this kind of description is accurate from modelling and the data point of view, it lacks the insight, particularly, in the presence of tens of similar expressions. Using fuzzy set, we can provide an intuitive and better expression as: IF the depart time is early morning and the travel time is long, THEN the travel mode is car-driving. Hence, we can combine multiple expressions into an easily understandable description and provide insight into the mode choice. Details of operations and algorithms of fuzzy sets are not included which can be found in (Klir and Yuan 1995).

Based on the features of fuzzy sets, we introduce several fuzzy attributes to replace some variables used in travel behaviour survey (see Section 3.2 for details).

3. CASE STUDY

3.1. Dataset description

The household travel survey (HTS) data is the largest and most comprehensive source of information on personal travel patterns for the Sydney Greater Metropolitan Area (GMA), which covers Sydney, the Illawarra Statistical Divisions and the Newcastle Statistical Subdivision. The data is collected through face to face interviews with approximately 3000-3500 households each year (out of 5000 households in the Sydney GMA randomly invited to participate in the survey). Details recorded include (but are not limited to) departure time, travel time, travel mode, purpose, origin and destination, of each of the trips that each person in a household makes over 24 hours on a representative day of the year. Socio-demographic attributes of households and individuals are also collected.

3.2. Fuzzy sets of travel mode choice variables

Based on the analysis of the character of exploratory variables of the dataset, we defined fuzzy sets for each of the two selected variables which are "depart_time" and "household_income".

In the survey dataset, the "depart_time" variable is recorded in minutes from 00:00 to 23:59 for the day. Following a Transport for NSW technical documentation (Bureau of Transport Statistics 2011), four fuzzy sets are defined for "depart_time" over the 24-hour period, which are "morning peak" (M), "evening peak" (E), "inter-peak" (L), and "evening/night period" (N). These fuzzy sets are illustrated in Figure 1.

In the survey dataset, the variable "household_income" indicates the annual approximate household income which ranges from –AU$5005.74 to AU$402741. Due to the spread of income, it is hard to get insight of the influence of the variable on travel mode choice. Hence, we introduced three fuzzy sets to depict easily understandable concepts which are consistent with people's ordinary experience on household income levels. The three fuzzy sets are "low income" (LI), "middle income" (MI), and "high income" (HI), which are shown in Figure 2, based on related information from the Australian Bureau of Statistics (ABS, Australian Bureau of Statistics 2012) and the Australian Taxation Office (ATO, Australian Taxation Office 2012).

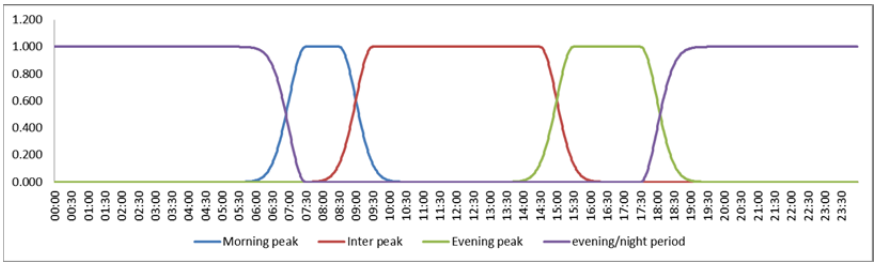

Figure 1: Fuzzy sets for "depart time"

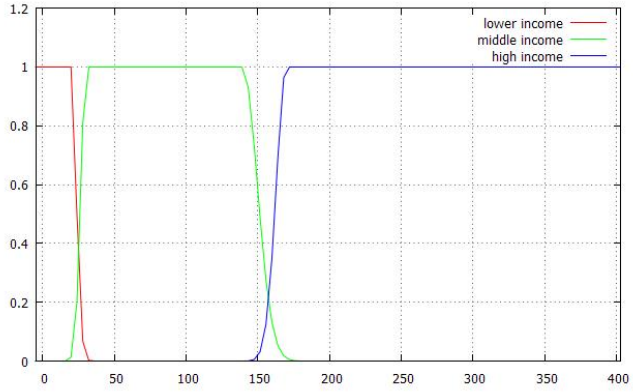

Figure 2: Fuzzy sets for "household income".

Following section discusses the results obtained from applying proposed methodology to the case study presented in this section.

4. RESULTS AND DISCUSSION

The presented method has been implemented and tested on a 100k sample which is randomly selected from a dataset for Sydney Household Travel Survey conducted by BTS, Transport for New South Wales (TfNSW), Australia. We partitioned the 100k sample into three subsets, i.e., a training dataset (30%), a testing dataset (35%) and a validation dataset (35%). The performance measure used for the comparison of classifiers is taken to be the percentage of records correctly identified (PCI).

Before conducting the test, we identified 17 variables based on statistical analysis of their correlations with the travel modes. These variables are trip_no, day_no, household_type, occupancy, veh_parked_here, hh_income, licence_num, student_sum, work_athome_sum, resident_num, pers_num_trips, trip_purpose, road_dist_xy, depart_time, arrive_time_tune, travel_mode, pre_mode. Among these variables, the variables "depart_time" and "household_income" are replaced by the fuzzy sets defined above when conducting the test for fuzzy settings. Further, we use previous trip's mode (pre_mode_new) and current trip mode (travel_mode) to test the scenarios when considering dependent/linked consecutive trips in a tour. Total 4 experiments (shown in Table 2) have been conducted based on different empirical settings (on DT and ANN) which are:

Experiment 1: We use travel_mode as decision variable and the others as explanatory variables. Under this setting, we test independent trip modelling and use this result as a benchmark for the following tests.

Experiment 2: Replacing the explanatory variables "hh_income" and "depart_time" by their fuzzy sets in Experiment 1. Under this setting, we test the performance of fuzzy sets in travel mode choice modelling.

Experiment 3: We add attribute "pre_mode_new" as an additional exploratory attribute to experiment 1 and test the performance of travel mode choice modelling based on linked trips.

Experiment 4: We add attribute "pre_mode_new" as an additional exploratory attribute to experiment 2 and test the performance of linked trips modelling based on consecutive trip under fuzzy set settings.

Table 2 gives the empirical settings and PCI of the eight experiments. Some observations from this table are:

A. Using dependent trips in a tour achieves higher PCI. For example, the PCI of experiment 1, 2 for both ANN and DT increases significantly from 64.71% to 85.63% in DT and 69.30% to 84.7% in ANN.

B. Using fuzzy sets as opposed to crisp

Table 2. Experiments Based on DT, ANN

Experiment	Empirical Settings		PCI (%)	
	Fuzzy sets	Dependent trip	DT	ANN
1	N	N	64.71	68.1
2	Y	N	67.67	68.7
3	N	Y	85.63	85.9
4	Y	Y	86.17	86.8

numbers gives higher PCI to ANN and DT. Experiments 1 & 2 for DT and 3 & 4 for ANN justify the use of Fuzzy sets. C. ANN performs better than the DT for all the experiments.

Based on the experiments, we can claim that our method can improve the PCI of travel mode choice. Table 3 illustrates the mode shares predicted by proposed approach considering ANN with fuzzy sets and tour based trips and it is compared with the original mode shares from HTS data. It illustrates that the mode shares from proposed approach are consistent with that from HTS data.

5. DISCUSSION AND CONCLUSIONS

This paper describes a novel methodology for travel mode choices based on data mining methods such as ANN and DTs combined with fuzzy sets. The proposed method considers (i) expert judgments by using fuzzy sets instead of crisp numbers for some explanatory variables; and, (ii) using the tour-based model that uses travel modes for previous trips as one of the predictor variables for current trip's mode choice. The proposed methodology is tested on a real dataset to evaluate the performance of classifiers for travel mode choice modelling. The results from various analysis conducted in this paper suggest that the use of fuzzy sets and tour-based model for mode choice achieves higher performances. In future, this work can be extended to include other explanatory variables, new fuzzy sets, and linking the individuals in the household to achieve higher classification performances.

Table 3. Mode Shares for ANN Prediction

Travel Modes	HTS data	DT Prediction	ANN Prediction
Car_driver	40.95%	43.50%	43.11%
Car_passenger	20.65%	30.76%	19.05%
Public_transport	8.37%	7.54%	7.74%
Walk	29.26%	17.68%	29.55%
Bicycle	0.77%	0.53%	0.53%

REFERENCES

Australian Bureau of Statistics (2012). Household Income and Income Distribution, Australia, 2009-2010. Canberra, Australia, Australian Bureau of Statistics.

Australian Taxation Office. (2012). "Parent, spouse's parent or individual relative tax offset calculator."

Australian Taxation Office. (2013). "Household Assistance Package - Tax Reforms."

Biagioni, J. P., & Szczurek, P. M. , Nelson P.C., Mohammadian K (2008). Tour-Based Mode Choice Modeling : Using An Ensemble of (Un-) Conditional Data-Mining Classifiers, (312).

Bureau of Transport Statistics (2011). Sydney Strategic Travel Model (STM): Modelling future travel patterns. Sydney, Transport for New South Wales Australia.

Cantarella, Giulio Erberto and de Luca, Stefano "Modeling Transportation Mode Choice through Artificial Neural Networks" Proceedings of the Fourth International Symposium on Uncertainty Modeling and Analysis (ISUMA'03) 2003.

Cirillo, C., Axhausen, K. W. (2002). Mode choice of complex tours : A panel analysis. Arbeitsberichte Verkehrs- und Raumplanung, Institut für Verkehrsplanung und Transportsysteme, ETH Zürich, Zürich., 142.

Dell'Orco, M, Circella, G & Sassanelli, D 2007, 'A hybrid approach to combine fuzziness and randomness in travel choice prediction', European Journal of Operational Research, vol. 185, 2007, pp. 648-658.

Edara, Praveen Kumar "Mode Choice Modeling Using Artificial Neural Networks" MS Thesis 2003

Hagan, M.T., and M. Menhaj, "Training feed-forward networks with the Marquardt algorithm," IEEE Transactions on Neural Networks, Vol. 5, No. 6, 1999, pp. 989–993, 1994.

Hensher, D.A. and Ton, T.T. "A comparison of the predictive potential of Artificial Neural Networks and Nested Logit models for commuter mode choice", Transp. Res. E, 2000, pp.155-172.

Khan, O, 2007, "Modelling Passenger Mode choice behaviour using computer aided stated preference data" PhD Thesis.

Klir, G. J. and B. Yuan (1995). Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice Hall.

Koppelman, F.S. and Wen, C.-H. Alternative Nested Logit Models: Structure, Properties and Estimation. Transportation Research Part B, Vol. 32, No. 5, 1998, pp. 289-298.

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, Volume 7, Page: 115 - 133.

McFadden, D. Conditional Logit Analysis of Qualitative Choice Behavior. Frontiers in Econometrics, P. Zarembka, ed., Academic Press, New York, NY, 1973.

Miller, E. J., Roorda, M. J., & Carrasco, J. A. (2005). A tour-based model of travel mode choice. Transportation, 32(4), 399–422. doi:10.1007/s11116-004-7962-3

Moller, Neural Networks, Vol. 6, 1993, pp. 525–533

Nijkamp, P., Reggiani, A., and Tritapepe, T. "Modelling interurban transport flows in Italy: a comparison between neural network analysis and Logit analysis", Transp. Res. 4C, 1998, pp.323-338.

Reggiani, A. and Tritapepe, T. "Neural networks and Logit models applied to commuters' mobility in the Metropolitan area of Milan", in Neural networks in Transport systems, Ashgate, 1998.

Shmueli, D., Salomon, I. and Shefer, D. "Neural network analysis of travel behavior: evaluating tools for prediction", Transp. Res. 4C, 1996, pp. 151-166.

Wu, X., Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, "Top 10 algorithms in data mining" Knowledge and Information Systems, (2008) Volume 14, Issue 1 , pp 1-37.

Xie, C., Lu, J., Parkany, E. "Work Travel Mode Choice Modeling Using Data Mining: Decision Trees And Neural Networks," Transportation Research Record: Journal of the Transportation Research Board, No. 1854. (2003)

Yaldi, G. (2005). Developing A Fuzzy-Neuro Travel Demand Model (Trip Distribution And Mode Choice), 1–15.

Zadeh, L. A. (1965). Fuzzy Sets, Information and Control, vol. 8, pp. 338-353.

Zhang, Yunlong and Xie, Yuanchang "Travel Mode Choice Modeling with Support Vector Machines" Transportation Research Record: Journal of the Transportation Research Board, pp 141-150, 2008