Data Mining Approach to the Work Trip Mode Choice Analysis in the Chicago Area

Yandan Lu
College of Urban Planning and Public Affairs, University of Illinois at Chicago
412 S. Peoria St., Chicago, IL 60607-7065
Phone: (312) 413 - 0746
FAX: (312) 413 - 0006
E-Mail: danalu2009@gmail.com

Kazuya Kawamura (corresponding)
College of Urban Planning and Public Affairs, University of Illinois, Chicago
412 S. Peoria St., Chicago, IL 60607-7065
Phone: (312) 413 - 1269
FAX: (312) 413 - 2314
E-Mail: kazuya@uic.edu

Word count: 5406 + 8 Tables and 1 Figures = 7656

March 15, 2010

Abstract:

Discrete Choice Methods have brought attention to the study of the travel mode choice behavior in both theoretical and practical areas of transportation planning and modeling. Unlike discrete choice models that impose pre-defined probability distributions on choice probabilities, data mining approaches view the travel mode choice as a pattern recognition problem whereby the travel choices can be identified by a combination of explanatory variables. As such, data mining techniques have enjoyed increasing applications in agent-based modeling. This paper examines the capability of a promising machine learning algorithm, Class Association Rules (CAR), for the estimation of the work trip modal choice using Chicago Area Transportation Studies (CATS) 1990 Household Travel Survey (HTTS) as the case. The purpose of the paper is to investigate the advantages, disadvantages and applicability of CARs for the study of mode choice behavior. It reveals that the CAR is a useful tool for building a powerful mode choice model, with the overall accuracy reaching up to 93% for the data set used in this study. Unlike some of other mining methods, the rules extracted by CAR are easy to interpret and provide insights into the travel behavior from a perspective that is different from statistical models. The example presented in this paper also illustrates one of the key advantages of CAR over discrete choice models, which is the flexibility of the model specification.

Key Words: Mode Choice, Data Mining, Travel behavior

INTRODUCTION

Discrete Choice Methods have brought attention to the study of travel mode choice behavior in both theoretical and practical areas of transportation planning and modeling. However, Random Utility Maximization theory, the principle behind most traditional discrete mode choice models, has been criticized for its inability to incorporate limitations in travelers' computational capability, information availability, and uncertainty. Unlike discrete choice models which impose pre-defined probability distributions on choice probabilities, data mining approaches view travel mode choice as a pattern-recognition problem whereby the travel patterns can be identified by a combination of explanatory variables, such as household characteristics, personal properties, and trip attributes. By exploring the data, the data mining approach can identify the patterns and connections between those explanatory variables and the choice of mode.

Data mining approaches have several advantages over discrete choice model. For example, Xie et al. (*1*) conducted a thorough comparison of the predictive performances of decision trees, neural network, and multinomial logit (MNL) models for mode choice. The study revealed that decision trees demonstrated higher estimation efficiency and better explicit interpretability. The trees recognized the patterns and connections among household characteristics, personal attributes, trip properties, and the choice of mode. Xie et al. also provided a scan of the studies that examined the application of neural network in transportation research.

In addition to those practical advantages that Xie et al. identified, machine learning algorithms in the data mining approach such as: decision tree, classification using association rules, and naïve Bayesian classification, are able to learn the classification from the data that can be used to predict the classes of new cases. Introducing the learning property to mode choice enables households to adapt to the environment and thus evolve. It is essential that households have such ability when simulating the interaction and dynamic behaviors over time. Third, the rules reflect the heterogeneous characteristics of the behaviors hidden in the data. Unlike discrete choice analysis that reveals the heterogeneity through the coefficients, the rules themselves reflect the heterogeneity. Finally, there is no concern for the correlation between explanatory variables in a data mining approach.

Lu et al. (*2*) applied a data mining approach, Class Association Rules (CARs), to study the mode choice behavior and then integrated the mined rules to an agent-based model (ABM) which simulated the effect of urban forms on work travel behavior. Their research exhibited another advantage of data mining approach when it is used in the ABM framework since the rule-based simulation of decision-making matches the essence of agent-based modeling. For example, RAMBLAS (*3*) and ILUTE (*4*) incorporate rule-based approach in land use - transportation interaction simulation. In addition, they share the same developmental roots in Artificial Intelligence.

Compared to the mode choice model used by Lu et al., the model presented in this paper includes two major improvements. One is that explanatory variables have been refined to better represent the situation in which travelers make mode choice decisions. For example, the attributes relating to all options of modes, not only the mode that was eventually chosen, are presented in the data set. The other is that a new method has been developed to interpret the mined rules. Even though the rules provide a straightforward relationship between explanatory variables and the mode chosen, the direction and strength of relationship are not quantitatively expressed. Most studies that use data mining techniques rely solely on predictive accuracies to evaluate the overall performance of rules (*1, 2*), but in the process incur tremendous loss of information by using such an aggregate measure. To our knowledge, no existing literature explores the techniques for analyzing the rules in detail to understand and obtain insights into the relationship between the explanatory variables and the choices made, in a manner that is analogous to the way discrete choice models are often used.

This paper examines the capability of a promising machine learning algorithm, Class Association Rules (CARs), for the estimation of work trip modal choice using Chicago Area Transportation Studies' (CATS) 1990 Household Travel Survey (HTTS) as the case. The paper also applies prediction accuracies to evaluate the performance of the models and also demonstrate the use and interpretation of the results. Finally, it demonstrates the advantages, disadvantages and applicability of CARs for the study of mode choice behavior over logit/probit models.

METHODOLOGY

Association Rules Mining

In this study, class association rules (CARs), a machine learning approach, is examined as the rule mining technique for mode choice behavior. The CARs integrates two mining approaches, association rules and classification rules, to mine a special subset of association rules (*5*). Association rule mining tries to extract the correlations, patterns, and

associations among item sets in a data repository (*6*). It is currently one of the most important and popular data mining techniques, and thus we are able to provide only a brief overview of some of the key works. It was originally applied to transaction databases to understand consumers' purchasing decision behavior (*7*). For example, a mined rule shows that consumers who usually buy eggs (e.g. 90%) tend to buy milk also. Later on, spatial association rule mining (*8*) and temporal association rule mining (*9*) were developed to provide more insights into people's decision behavior. Spatial association rule mining includes spatial information in the associations between one set of features and another. For example, a rule describing the relationship between homeownership, car ownership, and car use shows that home owners drive to work and the confidence of this rule is 70%. If spatial information on the trip origin and destination are available, the rules may conclude that home owners living near a train station and working in the city center take a train to work with a 90 % confidence. Temporal association rule mining adds time factors in the rule and thus can be even more informative. For example, home owners living near a train station and working in the city center take a train during peak period hours with a confidence of 95%. However, the confidence may drop to 20 % during off-peak period.

Class Association Rules (CARs)

Since association rule mining extracts all associations among variables in the data, it usually creates a large number of rules that are hard to interpret and implement. Thus, to create a usable set of rules for prediction of behaviors, researcher often focuses on a target behavior and identify other features or behaviors that are related to this target behavior. CARs mining achieves this purpose by extracting a small set of rules with predetermined targets (*10*). Other methods that use pruning (*11*) or cover rules (*12*) have been proposed to mine a small set of rules, but the results were generally not encouraging. Liu et al. (*5*) proposed an efficient algorithm, called Classification Based on Associations (CBA), to identify a subset of association rules that satisfy the predefined performance conditions, support and confidence, that can be tailored to specific applications. This process is performed using classification rules mining. They applied this algorithm to 26 databases which included: public health, image recognition, labor market, etc. The results showed that the rules derived from CARs were more accurate than those produced by the decision tree using C4.5, the most popular algorithm for decision tree mining.

There are four important concepts in CARs: rule item, support, confidence, and performance accuracy. The following gives a detailed description of these concepts when they are used in travel mode choice analysis.

Rule Item

A rule item always consists of a condition set, which is identified by a set of explanatory variables and their respective values, and a transportation mode, for example, a rule item (DriveTime < 4 min, Cars = 0-> Mode = walking) shows that if a trip takes less than 4 minutes by auto and the traveler has no car, then he will walk to this destination. More complicated rules contain more explanatory variables, especially when the rules are used to distinguish two similar modes such as driving and passenger in a car, or using a train and a bus. To evaluate the quality of a rule item, two more concepts, support and confidence, are introduced as below.

Support

Support of a rule item measures the popularity of a decision behavior among the population. It is defined as the ratio of the number of cases in the dataset that contain both the condition set and the mode over the total number of the cases in the dataset (*5*). So, if the support of a rule item is 50 % and the sample is representative, it means half of the travelers behave as described in this rule item. Thus the higher the support, the better the rule performs in capturing the travel behavior pattern. It is possible to set up a minimum support threshold exogenously before performing the data mining. If the support of a rule item is greater than the minimum support, the rule is assumed to be frequent. The minimum support is exogenously defined so that it is flexible to mine the rules that satisfy the requirement of frequency. Technically, the minimum support can be set to 0, which enables CARs to mine out all patterns of mode choice behavior hidden in the dataset even though just one person follows this pattern.

Confidence

Confidence of a rule item measures the popularity of a decision behavior among the population with similar attributes. It is defined as the ratio of the number of cases in the dataset that contain both the condition set and the mode over the

number of cases in the dataset that contain the condition set (*5*). If the confidence of a rule item is 90 %, it means among the people with the attributes defined by the condition set, 90 % choose the mode described by the rule item. Thus the higher the confidence, the better the rule item does in capturing the behavior of the homogeneous group. If the confidence of a rule item is greater than the minimum confidence, the rule is assumed to be accurate. The minimum support and confidence are exogenously defined so that it is flexible to mine the rules that satisfy the requirement of accuracy.

Performance accuracy

Once the rule set is derived, the application to predict the mode chosen for a traveler is very straightforward. It is exactly like an if-else process. A traveler goes through the rules one by one in order. Once his condition fits into one set of rules, the corresponding classifier is chosen as his travel mode and the process ends. Since the rules are sorted by confidence and then by support, the mode predicted by the rule set is always the one with the highest accuracy. Performance accuracy is used to measure the overall performance of the rule set to accurately predict each kind of travel mode. The higher the accuracy, the better the rule set is for identifying the mode. For example, if the performance accuracy for driving is 90 %, it means 90 % of drivers in the sample are correctly identified.

DATA

The Chicago Area Transportation Study (CATS) 1990 Household Travel Survey (HHTS) data (*13*) came from two data sources: Data collected locally and data collected by the Census Bureau. Basically, the census provided the journey-to-work trip information for the Chicago metropolitan area. The locally collected data provided the non-work related travel information. Research spanning seven years was conducted to produce a body of information on both work and non-work trips in HHTS. A total of 19,314 households responded from the six counties in the Chicago Area. Each respondent provided a 1-day travel diary and detailed individual and household socio-demographic data for the database. The whole database consists of three interrelated data sets reflecting household, person, and trip characteristics.

From the data set, only the trips that started at home and ended at the work place were considered. In the HHTS trip file, any chained trip from home to work was recorded in multiple segments. For example, a driver might need to pick up or drop off somebody on his way to work. In the HHTS, this trip was broken down into two segments: from home to pick up/drop off location, and from there to the work place. If a person took transit to work, the trip could be divided into more segments. For example, suppose that a person walked to a bus station, took the bus to a train station, then took a train to the destination station. From there, he walked to the work place. His trip would thus be divided into four segments, each consisting one record in the HHTS trip file. Before these chained trips could be used for CARs, all the segments had to be aggregated into one record so that a single record contained all the information for a chained journey from home to work place.

Four travel modes were considered in the development of the mode choice model. They were: driving, passenger in a car, train, and walking. For driving trips, the data were further simplified. Only the non-stop trips from home to work place were used. For transit riders, there were usually several trip segments. The segment on a train, including the two rail systems in the region, CTA and METRA, was treated as the primary segment. All the trip segments occurred before or after the primary segment were treated as access trips. A summary of the information retained for each mode is exhibited in Table 1. As shown in this table, even though each mode has total travel time, the time occurs when that mode is used.

TABLE 1 Available information for each mode

The HHTS data is a revealed-preference data set in which the information is provided only for the actually travel decisions that materialized, but not for the alternatives that might have been considered but were not chosen. For example, the data on a driving trip contains information related to the driving option such as driving time, but no information is provided for the possible alternative travel choice using a train. Just as in the discrete choice models, in order to better represent the situation a traveler faces when the mode choice decision is made, the variables for other options need to be estimated for each record in advance.

The data used to derive alternative travel modes' attributes include the output data from the Chicago Metropolitan Agency for Planning (CMAP) travel demand model containing the estimated driving time between each

travel analysis zone (TAZ) pair, the data mapping CATS quarter sections (i.e. 0.5 mile by 0.5 mile grids) to TAZs, train station GIS shape files, and the databank for the CMAP model which holds the data for performing the trip assignment for the Chicago urban area. CMAP is the metropolitan planning organization for the Chicago region, and the successor to CATS.

For the train trips, driving time from one's home to work place was estimated by the user-equilibrium travel time from the CMAP model output at the TAZ level. The trip file in the HTTS data had spatial information for trip origin and destination, recorded at the quarter section level for trip origin and also destination. The quarter sections were then mapped to TAZ so that the estimated driving time from the model could be merged to the HHTS data for train riders.

For each driving trip, the attributes of the train option needed to be estimated. These attributes were identified by three variables. They were: the distance from home to the nearest train station, the distance from the destination station to the work place, and travel time using the train including transfers. The CATS quarter section GIS file and the train station GIS file were used to derive the two access distances. The HHTS data provided the CATS quarter section information of the origin and destination of a trip. The distance from the centroid of a quarter section to its nearest train station is used to estimate the access distances.

The third variable for the train option, the travel time using the train from the origin train station to the destination train station, was estimated by the output from the CMAP model. It provided the travel times between two neighboring stations by train, which are essentially the link travel times. Since it is quite common that multiple lines use the same station, especially in the Loop, there were multiple travel times for the same pair of adjacent stations. In such a situation, they were examined one by one manually to select the most reasonable travel time. To derive the in-vehicle travel time between each pair of train stations, a small program was developed to calculate the shortest path between two stations, in terms of travel time, using the Dijkstra's algorithm (*14*). The shortest path travel times between all possible station pairs, considered as the in-vehicle travel time by train, were then merged to the HHTS.

Using the methods described above, each trip in the HHTS was provided the travel attributes of the alternative modes that were not actually chosen. The data processing and merging for walking trips, which require information for both driving and rail alternatives, exactly followed the approach described above for driving and train trips.

The trip file was then merged to the household-person file. The resulting file contained all the household properties, personal characteristics, and trip information. The total number of cases was 9,210 with 222 walking trips, 7961 driving trips, 408 passenger trips, and 619 train trips.

MODEL SPECIFICATIONS

Four modes are explicitly represented in this model. They are: walking, driving, passenger in a car, and train. Three sets of explanatory variables, as shown in Table 2, are assumed to have significant effect on people's work travel mode choice behavior. They are (a) household properties (b) personal attributes and (c) trip characteristics.

Household properties include: household income, which is divided into seven cohorts, household size, which ranges from 0 to 10, and the number of vehicles in the household, which includes auto, pickup truck, and motorcycle and takes value as 0 or 1+. Personal attributes consist of age and gender. Only those persons whose ages fall into the range from 18 to 70 are taken into consideration as decision makers since this study focuses on the work trip mode choice. A work trip refers to the whole trip in a home-to-work activity chain. For the driving trips, this study only concerns the trips starting from home and ending at a work place in one trip segment. Those driving trips with other travel purposes on the way to work, i.e. chained trips, are not considered. For the mode choice, only driving time and the vehicle operating cost are considered. The vehicle operating cost is derived by multiplying the driving distance and driving cost per mile. Driving cost per mile covers direct expenses, such as the insurance, depreciation, fuel, maintenance, parking, and travel time. It is exogenously defined as 0.8735 dollar per mile (*15*), of which 6 cents per mile are for the fuel cost. It turns out that in terms of mode choice predictions, the cost of driving per mile has no effect on the output as long as it is linear with driving distance because the mined rules simply adjust the cost threshold according to the cost of driving. Transit trips are divided into three segments: the access trip from home to the origin train station, travel on the train (i.e. line-haul), and the access trip from the destination train station to the work place. In the mode choice model, the two access trips are measured in terms of distances and the travel on the train is measured by in-vehicle travel time, as shown in Figure 1.

TABLE 2 Explanatory variables for mode choice

FIGURE 1 Segments of a train trip

For the development of the mode choice rules, a total of 9,210 observations from the HHTS were used. Only the morning peak period work trips in the data set were used. To mine the mode choice rules, the data were randomly split into two groups. One contained 80 % of the data for training the model, and the other 20% for testing. Table 3 provides the mode split information for the two groups of data.

INTERPRETATION OF THE RESULTS

While data mining is often regarded as a rather mechanical exercise that identifies patterns from a sea of numbers, the rule set produced by CARS offers insights about the relationship between the variables and the outcome. To mine all the rules hidden in the data, the minimum support and minimum confidence are set to 0. A total of 637 rules were mined. Each rule contains three parts: condition set, classifier, and parameters for support and confidence. Rules are sorted by confidence first and then by support.

TABLE 3 Mode splits for the data set

Performance Accuracy

Table 4A shows the results for applying the rules to the training data; while Table 4B represents the results when the rules are applied to the testing data. The overall accuracy of predicting correct mode is 96.05 % and 93.37 % for the training data and the testing data, respectively. As shown in the two tables, for the training data, the rules performed very well for predicting walking, driving and train modes. Their accuracies are 96.24%, 99.89%, and 97.77%, respectively. When the rules are applied to the testing data, accuracies dropped to 63.89%, 98.93%, and 85.71%, respectively. While the model performs adequately in predicting the minority modes such as rail and walking, it does a poor job of predicting the travelers who are passengers of automobiles, with the accuracies of just 19.82% and 2.67% for the training data and the testing data, respectively. In most cases, the model is not able to distinguish between the passenger and driving, and the majority of the former are predicted as the latter. It also reveals that drivers and passengers in car have very similar attributes and behavior patterns. The difference between these two modes cannot be well identified by the explanatory variables considered in this model.

TABLE 4 Test results of mode choice rules

Of the 637 rules, 462 have 100% confidence and 154 have confidences between 90% and 99%. Hence about 97% of the rules have confidences higher than 90%. This explains the high predictive accuracy of the model. The highest support of 5.7% is observed for the rule with a confidence of 95.81%. The highest support value for the rules with 100% confidence is just 2.94%, equal to 271 observations. The low support levels indicate that mode choice is a very heterogeneous and complicated behavior. Even though all rules have been mined out, no obvious pattern that explains the behavior of the majority exists. It should be noted that the rules are derived only for the work trips that can be considered as relatively consistent in terms of mode choice decision factors. It is reasonable to assume that the mode choice behaviors for other purpose trips are even more complicated and diverse.

Analysis of the Mode Choice Model

Of the 637 rules, 586 define conditions for driving and passenger modes, 118 for rail travel, and 33 for walking. Thus, driving and passenger modes dominate in term of the number of rules. The rules indicate that trip attributes, such as the distance from the destination station to the work place, the distance from home to the origin station, driving costs, and travel time on train, have strong effects on the mode choice behavior. Socioeconomic variables including car ownership and household income also play a significant role.

Table 5 shows the usage frequency for each of the variable in the rule set categorized by mode. Frequencies are weighted by the supports of the rules in which the variable is present. For example, variable INVEHTIME is used in Rule 1 whose support is 271. It means 271 observations use variable INVEHTIME in their mode choice decisions along with other variables included in Rule 1. Therefore, the frequency for this variable is 271. If INVEHTIME is also

used in Rule 2, the frequency will be accumulated. The frequency can be used to measure the importance each variable has on the mode choice behavior because the higher the frequency, the more often this variable is used by commuters in their mode choice decision.

TABLE 5 Usage frequencies of explanatory variables in the mode choice model

In general, the table shows that GENDER, AGE, ORGCBD (1 if the trip origin is in CBD), and DESTCBD (1 if the work place is in CBD) are the four variables that are least used in the rule set. In other words, they are not as significant as other variables for commuters' mode choice behavior. INVEHTIME is also used infrequently, but it is one of the critical variables in determining transit mode choice. For driving, the most important variables are household size, driving cost, the distance from home to the nearest train station, and the distance from the work place to the nearest train station. For the rail mode, driving time, time spent on the train, driving cost, and the distance from the work place to the nearest train station play more important roles. For walking, the driving time, driving cost, availability of vehicle, and living in CBD are critical.

To examine the effects of the aforementioned four key variables for rail travel in more detail, further analysis was performed. Table 6 provides the usage frequency by the range of values for each of the four variables. The value ranges with 0 usage frequency are not shown in the table. Table 6A shows that 1,041 train riders have driving times longer than 30 minutes, while 319 have driving times less than 30 minutes. It means that long-distance commuters are more likely to choose the train as the mode to work. This makes train a competitive mode against driving for the long distance trips in this model. Table 6B shows that 1,125 observations have the distance from the work place to the nearest train station less than 0.1328 mile, which is around 5 minutes of walking. But the access distance at the home end can be longer for the rail users. As shown in Table 6C, 0.91 mile, which is around 20 minutes of walking, is still acceptable for the rail users. Essentially, the mode choice rules suggest that the commuters would accept longer access distance for the home end than for the work end. The reason may be that there is a greater flexibility for accessing the station at the home end, e.g. kiss-and-ride, or drive-and-park, than for the work end. These rules also imply the greater importance of the spatial distribution of jobs on mode choice behavior. If there are more jobs located within the walking distance of train stations, the mined rules suggest that greater number of commuters would consider rail as an option.

Vehicle operating cost is calculated based on the aerial distance and the operating cost per mile. The ranges of DRIVECOST shown in Table 6D suggest that the vehicle operating of $13.7, which is equivalent of 15.7 miles, is the threshold for the rail to be competitive. Thus both Table 6D and 6A indicates that train is competitive for longer distance commute, especially when road congestion exists.

TABLE6 Usage frequencies of key variables for rail trip

Table 7 provides the usage frequency for driving trips by the range of values for each of the four variables: PERSONS, ACCMILE, DRIVECOST, and LEAVEMILE. Table 7B shows that around 90.4% of the drivers live at least 0.928 miles, around 20 minutes on foot, away from the nearest train station. Thus, about 9.6% of the drivers live within the walking distance from a train station but still choose to drive, probably because of poor train station access at the destination ends. Table 7C shows that around 93% of the drivers travel only moderate distances, 5 to 26 miles, to work. For excessively long trips (i.e longer than 26.4 miles), people prefer riding trains over driving. There are about 7% of the drivers who use cars to make short distance commute trips of less than 2.4 miles. Table 7D exhibits the same pattern as the rail trip, meaning that the travelers are more sensitive to the distance from the work place to the nearest train station than the distance from home to the nearest train station. This table shows that 23% of the drivers can access on foot from the work place to the train station in between 10 to 20 minutes. However, only 7% of the drivers live within 10 to 20minutes of walking to the nearest trains station. Again, this suggests that jobs need to be more clustered around train stations than homes in order to attract greater number of people to take trains.

TABLE 7 Usage frequencies of key variables for driving trip

Table 8 provides the usage frequency for walking trips by the range of values for the four most frequent variables: DRIVETIME, VEHS, DRIVECOST, and ORGCBD. The information conveyed by the four tables are straightforward. They show that a majority of walkers do not own a car. Also all of them live in the CBD. Furthermore, none of them lives more than 1.87 miles away from their work place.

TABLE 8 Usage frequencies of key variables for walking trip

6. CONCLUSIONS

For the last several decades, investigation of decision making behavior has been dominated by discrete choice analysis in social science. But the development of Artificial Intelligence in computer science brings another promising option. This research shows that the Classification Association Rules (CARs) is a useful tool for building a powerful mode choice model. It is very accurate, with accuracy reaching up to 93% for the data used in this study. In addition to its accuracy, the mined rules are easy to interpret and provide insights into the mode choice decision process. Also, as demonstrated in this paper, the analysis of rule sets generated by CARs reveals the patterns and idiosyncrasies in travel behavior that may not be easily obtained from statistical models. Also various forms of interactions among the variables, e.g. suppressing, intervening, in influencing the mode choice behavior can be revealed by the CARs output. The example presented in this paper illustrates one of the key advantages of CARs over discrete choice models, which is the flexibility of model specification. For example, it would have been challenging to simultaneously examine the effect of highly correlated variables such as driving time and cost.

However, there are several weaknesses that must be addressed. First, some variables are parsed too finely. For example, driving time, a continuous variable in the mode choice rules, is divided into 37 categories. This makes many categories too narrow, often just 2 minutes. The effects of such fine discretization of driving time on the sensitivity of mode choice behavior have not been investigated. Second, CARs, like decision trees, is a deterministic rather than stochastic method. It is worthwhile to investigate the applicability of other data mining approaches, such as Bayesian classifiers, which are stochastic. Third, CARs lacks theoretical foothold, unlike the very strong connection between the random utility theory and the discrete choice model. It makes CARs less than ideal for applications within the main-steam travel demand analysis framework at least in the eyes of many transportation professionals. Finally, unlike statistical models, the rules from data mining models do not provide quantitative interpretation of the effects of independent variables on the final decision-making. For example, the rules mined by CARs provide only the condition set and the final classifier. Therefore, the quantitative relationship between the variables in the condition set and the final classifier is not clear.

Overall, our analysis have shown that data mining is a very promising technique for performing mode choice analysis and deserves more research efforts to explore its values to the more broad transportation planning area. From practical point of view, the predictive performance of CARs shown in this paper may motivate practitioners to investigate this technique. For researchers, we believe its attraction is mostly academic at this point. Using techniques like CARs, one can simulate travel behaviors entirely based on rules instead of statistical distributions. That makes CARs better suited, at least theoretically, to the investigation of urban system from the complexity-based perspective.

ACKNOWLEDGEMENTS

This research was supported by the Lincoln Institute of Land Policy, College of Urban Planning and Public Affairs at the University of Illinois, Chicago. All responsibility for the contents of the paper lies with the authors.

References

1. Xie, C., J. Lu, and E. Parkany. Work travel mode choice modeling with data mining: decision trees and neural networks. *Transportation Research Record*, No 1854, 2003, pp 50-61.
2. Lu, Y., K. Kawamura, M. Zellner. The Influence of Urban Forms on Work Travel Behavior: an Exploration Using Agent-Based Modeling. *Transportation Research Record*. No. 2082. 2008, pp 132-140.
3. Veldhuisen, K. Jan, Harry J. P. Timmermans, and Loek L. Kapoen. 2000. Ramblas: A regional planning model based on the microsimulation of daily travel patterns. *Environment & Planning A* 32(3): 427-443.)
4. Salvani, P., and E. J. Miller. 2005. ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and Spatial Economics* 5(2): 217-34.
5. Liu, B., W. Hsu, and Y. Ma. Integrating classification and association rule mining. Paper presented at KDD-98, New York. 1998.
6. Kotsiantis, S., D. Kanellopoulos. Association Rules Mining: A Recent Overview. GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82

7. Agrawal, R., T. Imielinski, and A. Swami: Mining Association Rules Between Sets of Items in Large Databases". ACM SIGMOD Conference Record. Vol. 22: 2 1993: pp. 207-216
8. Koperski, K. and J. Han. "Discovery of Spatial Association Rules in Geographic Information Databases". Lecture Notes in Computer Science, 1995. Springer
9. Ale, J.M. and G. Rossi. "An Approach to Discovering Temporal Association Rules". Proceedings of the 2000 ACM Symposium on Applied Computing. 2000.
10. Sun, X. and Z. Wang, "An Efficient MA-Based Classification Rule Mining Algorithm," CSSE, vol. 1, pp.702-705, 2008 International Conference on Computer Science and Software Engineering, 2008
11. Schlimmer, J 1993. "Efficiently inducing determinations: a complete and systematic search algorithm that uses optimal pruning". ICML-93, 268-275.
12. Quinlan, R. and Cameron-Jones, M. 1995. Oversearching and layered search in empirical learning. IJCAI-95.
13. Chicago Area Transportation Study. CATS Working Paper 94-05: CATS 1990 Household Travel Survey - A Methodological Overview. Chicago Area Transportation Study. 1994.
14. Dijkstra, E.W. “A note on two problems in connexion with graphs”. In: Numerische Mathematik. 1 (1959), S. 269–271
15. Commute Solutions. The true cost of driving. http://www.commutesolutions.org/calc.htm. Accessed Jul 15, 2008

LIST OF FIGURES

LIST OF TABLES

TABLE 1 Available information for each mode

	Driving	Passenger	Train	Walking
Total travel time	Y	Y	Y	Y
Travel distance	Y	Y	Y	
Living in CBD	Y	Y	Y	Y
Working in CBD	Y	Y	Y	Y
Travel time in train			Y	

TABLE 2 Explanatory variables for mode choice

Variable	Definition	Values
HINC	Household income	1:less than $15,000; 2: $15,000 to $24,999; 3: $25,000 to $39,999; 4: $40,000 to $59,999; 5: $60,000 to $74,999; 6: $75,000 to $99,999; 7: more than $100,000; 0: unknown
PERSONS	Household size	Continuous
VEHS	Number of vehicles in the household	0: no car, 1: 1 car, 2: 2 or more than 2 cars
AGE	Age	Continuous
GENDER	Gender	1: Male; 0: Female
DRIVINGTIME	Travel time if driving	Continuous (minutes)
DRIVECOST	Travel cost if driving	Continuous (dollars)
ACCMILE	Distance from home to the nearest station	Continuous (miles)
LEAVEMILE	Distance from the nearest station to work place	Continuous (miles)
INVEHTIME	Travel time in train	Continuous (minutes)
ORGCBD	Dummy for living in the CBD	1: yes; 0: no
DESTCBD	Dummy for work in the CBD	1: yes; 0: no

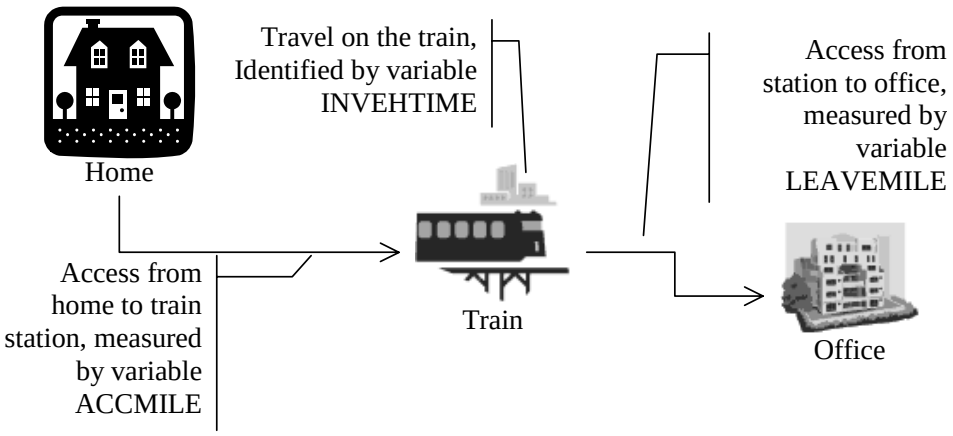

FIGURE 1 SEGMENTS OF A TRAIN TRIP

TABLE 3 Mode splits for the data set

	Total database		Training dataset		Test dataset	
Mode	Number	%	Number	%	Number	%
Walking	222	2.41	171	2.31	51	2.81
Driving	7,961	86.43	6,396	86.53	1,565	86.08
Passenger	408	4.43	331	4.48	77	4.24
Train	619	6.72	494	6.68	125	6.88
Sum	9,210	100	7,392	100	1,818	100

TABLE 4 Test results of mode choice rules

		Predicted mode				
		Walking	Driving	Passenger	Train	%Accuracy
Observed mode	Walking	179	7	0	0	96.24%
	Driving	2	6,365	4	1	99.89%
	Passenger	0	266	66	1	19.82%
	Train	0	10	1	482	97.77%

A. Prediction results on training data (80%)

		Predicted mode				
		Walking	Driving	Passenger	Train	%Accuracy
Observed mode	Walking	23	12	0	1	63.89%
	Driving	2	1,572	9	6	98.93%
	Passenger	0	72	2	1	2.67%
	Train	0	18	0	108	85.71%

B. Prediction results on testing data (20%)

TABLE 5 Usage frequencies of explanatory variables in the mode choice model

Explanatory variable	All rules		Rules for driving and passenger		Rules for transit		Rules for walking	
	Frequency	Order	Frequency	Order	Frequency	Order	Frequency	Order
PERSONS	15,276	1	14,850	1	373	7	53	7
DRIVECOST	15,021	2	13,909	3	930	4	182	3
LEAVEMILE	14,977	3	13,775	4	1,166	2	36	9
ACCMILE	14,588	4	14,182	2	384	6	22	11
HINC	14,085	5	13,731	5	307	9	47	8
DRIVETIME	14,081	6	12,472	7	1,360	1	249	1
VEHS	13,627	7	12,981	6	441	5	205	2
GENDER	12,635	8	12,249	8	321	8	65	5
INVEHTIME	12,605	9	11,447	9	1,094	3	64	6
AGE	5,271	10	5,240	10	31	11	0	12
ORGCBD	357	11	204	11	37	10	116	4
DESTCBD	188	12	153	12	0	12	35	10

TABLE6 Usage frequencies of key variables for rail trip

Value range	Frequency	Subtotal
60 – 64.6	90	1041
55 – 59.95	83	
50 – 54.95	128	
45 – 49.95	110	
42 – 44.9	85	
40 – 41. 95	54	
36 – 39.95	163	
35 – 35.95	17	
30 – 34.9	311	
25 – 29.95	76	319
20 – 24.8	110	
15.5 – 19.9	112	
13 – 14.8	12	
7 – 7.9	5	
6 – 6.9	2	
2 – 4.05	2	

A. Variable: drivetime (unit: minute)

Value range	Frequency
0 – 0.1328	1125
0.264 – 0.268	27
0.288 – 0.293	8
0.293 – 0.45	6

B. Variable: leavemile (unit: mile)

Value range	Frequency
0.11 – 0.42	119
0.42 – 0.91	211
0.928 – 1.4	20
> 1.4	34

C. Variable:accmile (unit: mile)

Value range	Frequency
> 23	366
13.6 – 23.1	453
4.4 – 13.6	97
2.11 – 4.4	14

D. Variable: drivecost (unit: $)

TABLE 7 Usage frequencies of key variables for driving trip

Value range	Frequency
1	3429
2	4977
3	3020
4	2316
5	85

A. Variable: PERSONS

Value range	Frequency
0.109 – 0.421	265
0.421 – 0.907	1100
0.928 – 1.404	1473
> 1.404	11319

B. Variable: ACCMILE (unit: mile)

Value range	Frequency
0 – 0.807	44
0.807 – 1.635	727
1.635 – 2.11	187
2.11 – 4.414	2670
4.414 – 13.657	8606
13.657 – 23.068	1595
> 23.068	60

C. Variable:DRIVECOST (unit: $)

Value range	Frequency
0.131 – 0.264	300
0.268 – 0.287	10
0.292 – 0.458	798
0.458 – 0.908	2405
0.925 – 1.403	1923
> 1.419	8324

D. Variable: LEAVEMILE (unit: mile)

TABLE 8 Usage frequencies of key variables for walking trip

Value range	Frequency
0 – 0.5	88
2.05 – 4.05	53
4.05 – 4.95	32
5.05 – 6.05	18
6.05 – 6.9	39
7.05 – 7.9	19

A. Variable: DRIVETIME (unit: mile)

Value range	Frequency
0	192
1	13

B. Variable: VEHS

Value range	Frequency
0 – 0.807	91
0.807 – 1.635	91

C. Variable:DRIVECOST (unit: $)

Value range	Frequency
1	116

D. Variable: ORGCBD