An Innovative System of Disaggregate Models for Trip Generation and Trip Attributes Using Random Forest Concept

Milad Ghasri, Taha Hossein Rashidi

Abstract

This research attempts to address the gap between research and practical transport demand models. Evolution of travel demand models started from the early paper and pencil versions of conventional four-step models of the late 50s and proceeded to the activity-based models.

During this transition the emphasis shifted from aggregate to disaggregate models, whereby researchers increasingly paid attention to individual decision making regarding daily activities. Meanwhile, the role of computational improvements, which prepared the ability of manipulating large datasets, is undeniable. This alteration engendered a conspicuous disparity between aggregate and disaggregate models regarding their practicality, precision and policy sensitivity level. This disparity or trade-off results in a gap in the state-of-the-art and practice of travel demand modelling which is the main attention of this paper. Unlike the four-step models, a highly disaggregate travel demand modelling structure is proposed in this paper in which trip purpose, mode of transport, time of day, commute distance and destination choice decisions are modelled. All these decisions are jointly modelled using the random forest method which is an advanced data mining method. Predicting total number of generated trips is the first step in the proposed framework and first trips' attributes is then sequentially modelled. A serial correlation is considered among consequent trips in order to jointly model travel attributes. The proposed system of models discloses outstanding accurate results for the application developed using the 2007 wave of Victorian Integrated Survey of Travel and Activity (VISTA).

Introduction

Developing an accurate and efficient travel demand model has been a challenge for travel modellers since the introduction of traditional four step models. Several criticisms, as policy insensitiveness, are associated with the aggregate travel demand models, although they have been employed for many metropolitan areas around the world. With the intention of tackle aggregate models' drawbacks, researchers get more interested in disaggregate approaches. As both data and computational access have grown, benefits of disaggregate models are more comprehended for planning. Disaggregate models are developed in a bottom-up way therefore the actual behaviour of individuals can be observed and because of this behavioural characteristic they are qualified to capture the impact of policies on people's decisions. Since the introduction of activity-based models (1) several applications of them have been developed. At the same time the conventional and popular four step models are being less studied and developed by researchers and practitioners due to their policy sensitivity restrictions (2). Nonetheless, the cost, including time and money, of developing the highly disaggregate activity-based models is considerably higher than developing the aggregate models (3). As a result, especially for small and medium sized cities, the need for a disaggregate travel demand model which is not costly as an activity-based model appears to be essential.

Most critically, while disaggregate models have substantial appeal, considerable research is required to further understand them and employ them towards planning purposes. Despite advances in computational capabilities, still one of the most essential issues associated with disaggregate models is the cost of running such models. This research attempts to introduce a framework which not only is developed in disaggregate approach, but also is computationally fast. This new approach introduces a cohesive structure to simulate several travel demand variables that are required for a traffic assignment model. Such a disaggregate travel demand model should replace the trip generation, trip distribution and mode choice steps of an aggregate four-step model while it is not as aggregate as four-step models. In other words, it is essential that mode choice, trip purpose, time of day and trip distance models would be jointly modelled as these decisions have a high inter-dependency. Nevertheless, destination choice model still needs further attempts to be fully functional. Once coupled with the destination choice model, the complete framework can substitute a four-step model while being even less complicated and being policy sensitive like an activity-based model.

Being mostly categorical with large possible levels, the dependent and independent variables of the proposed framework demand a particular modelling, and data mining approaches found to be appropriate to answer this characteristic. Random Forest (RF), one of the advanced data mining's methods, is an efficient method for dealing with classification problems (4). Besides, the concept of out of bag (OOB) in this method, enrich us to check the accuracy of the models internally.

Methodology

The modelling presented in this part is envisioned to be able to substitute a four-step modelling system. The framework, displayed in Figure 1, starts with a trip generation model which is known to be the most transferable travel attribute (5). Once total number of trips per day is estimated, first trip's attributes would be determined utilizing model 2-1 for those people with nonzero number of trips. Trip attributes includes trip purpose and its starting time. In the next step model 2-2 would predict destination attributes including commute distance, land use variables of destination and mode of travel. It should be emphasized that in each step, outputs of the previous models are considered as input, or independent variables, for the current model. Models 3-1 and 3-2 are respectively the same as models 2-1 and 2-2; however they are used for next trips of individuals and the pre-estimated attributes of previous trips could be considered as explanatory variables for them. There is no need to say that models 3-1 and 3-2 are utilized just for those with more than one trip per day.

Although presented framework is a comprehensive system of models that is able to predict trip diary for individuals, its output still cannot be used for assignment. The reason is that models 2-2 and 3-2 are defined to predict destination attributes not their exact location. In other words, the main problem of estimating travel demand matrix is separated into two sub-problems: estimating individuals' trip diary based on quantitative and qualitative attributes of destination, and spotting exact location of destination based on estimated attributes. Current study addresses the first sub-problem which is more complicated and commute distance which is one of the framework's outputs is estimated to facilitate the destination choice problem through decreasing the choice set of destinations.

The latest data transferability (6) study used the decision tree method to model some travel attributes in a sequential paradigm. Although the general goodness-of-fit of that study was acceptable, a more advanced method of random forest is employed in this paper as it is required to have higher accuracy due to the complexity of the models developed in this paper.

In random forests, no cross-validation or a separate test set is required to get an unbiased estimate as it is estimated internally, during the run. Assuming that the reader is familiar with tree classification methods, each tree in the RF is constructed using a subset of data which is bootstrapped from the original data with replacement. The portion of data not included in the tree development is called out-of-bag (OOB). The accuracy of the forest is estimated using the OOB as a test set for each tree. Once the forest is constructed, each case has been left out several times depending on the number of trees. Those trees for which the case observation is in the OOB set are then considered for classification and estimation of the OOB error rate. Using the OOB error rate, RF provides an internally estimated unbiased error estimate which has benefits of the cross classification testing approach.

Travel attribute modelling exercise of this paper can be distinguished from the previous methods in several directions. First, the travel attributes are divided in two category and each of them is jointly modelled which was the major missing issue in the past. It is not only the travel attributes that are jointly modelled, the impact of the attributes of the previous trip are regarded when travel attributes are simulated in the current study. This is also an essential advantage of the proposed framework as it moves the system from a static phase to a dynamic one. Employing RF as an advanced data mining method is another advantage of the proposed system of models which a flexible platform that enhanced the accuracy of the proposed system of models. The framework presented in Figure 1 benefits of having an internally linked destination choice model which was also missing in the previously developed data transferability models.

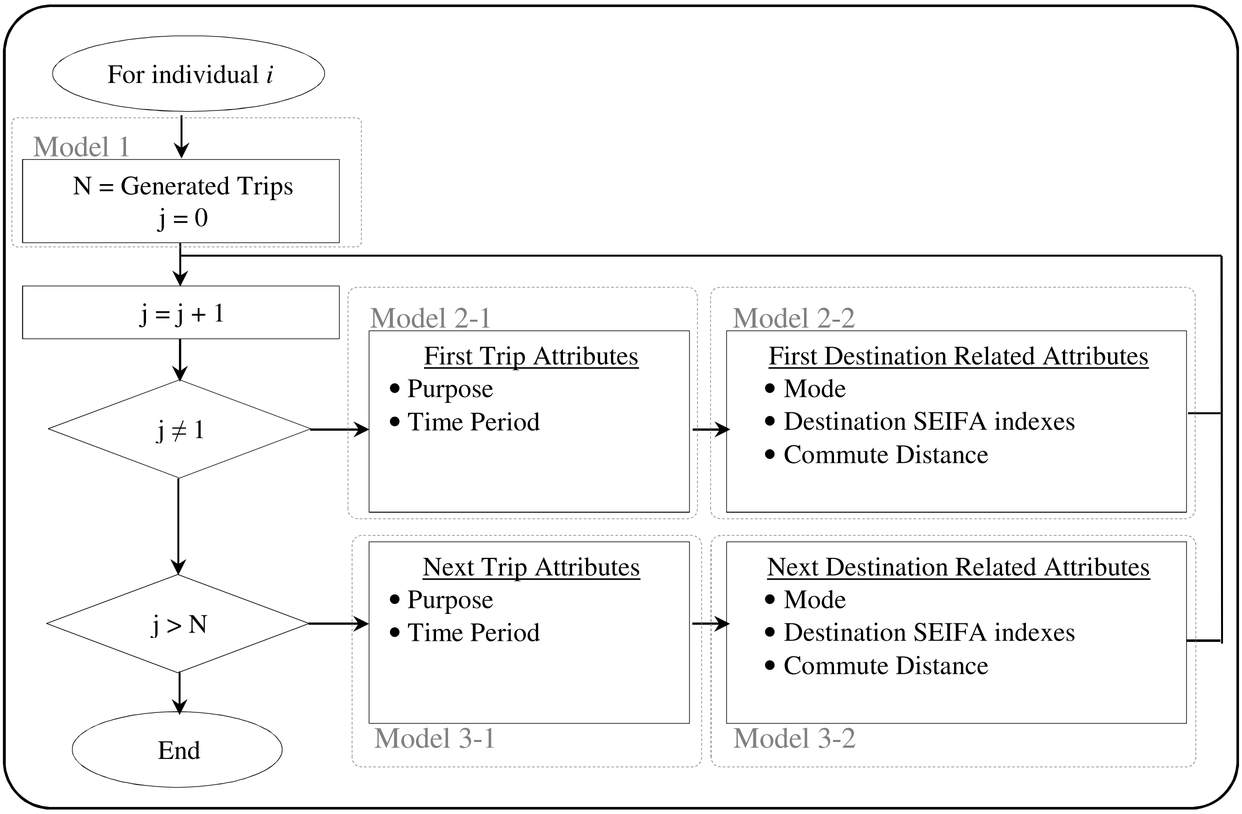

Figure 1 – General flowchart for the modelling structure of the study

Data

The main data sources used in the study are the household travel surveys collected for major Australian cities of Sydney, Melbourne and Brisbane (7, 8 and 9). Daily trips in Melbourne for the 2007-2008 VISTA were used. After cleaning the data, 76,015 trips for 25,484 individuals were kept in the final dataset. The cleaned data includes trip purposes of work, education, getting service of buying goods, pick up of drop off, recreation and return for three modes of transport of auto, transit and non-motorised in five duration of before morning peak, morning peak, between morning and afternoon peak, afternoon peak and after afternoon peak, and in four commute distance categories of less than two kilometres, between two and four kilometres, four and eight kilometres and greater than eight kilometres.

For different modelling purposes of this paper, in total, 33 explanatory variables were included. These explanatory variables can be grouped into two major groups: socio and economic -demographic attributes and built environment variables. Socio and economic –demographic attributes used in the data include sex, age, availability of driver's license, occupation attributes such as type and being full time or part time, household size, number of cars, and income.

The built environment variables are extracted from the 2006 Socio-Economic Indexes for Areas (SEIFA) data. SEIFA is a product developed by the Australia Bureau of Statistics that ranks areas in Australia according to relative socio-economic advantage and disadvantage. The 1) Index of Relative Socio-Economic Disadvantage (IRSD), 2) Index of Relative Socio-Economic Advantage and Disadvantage (IRSAD), 3) Index of Education and Occupation (IEO), 4) Index of Economic Resources (IER) provided in SEIFA are based on information from the five-yearly Census. The census collection districts (CD), in which the home and the origin of the trip are located, are used to be joined to the corresponding CD in the SEIFA data. House location's SEIFA indexes are treated as continues independent variables for all the five models. However, since whatever is used as origin related explanatory variable for trip attribute and destination attribute models should be estimated as the destination attribute for the previous trip, each of the trip origin's SEIFA indexes is classified in three major groups and then used as independent variable in trip attribute and destination attribute models.

Maximum generated trip for kept individual is 19 and minimum value for that is 0 which leads to 20 levels for this variable. Trip attributes include trip purpose and trip starting time, by definition, and since they have 6 and 5 levels respectively, their combination results in 30 possible levels.

For destination attributes model transport mode, commute distance and classified SEIFA indexes should be estimated that brings up 219 levels for first trips (model 2-2) and 238 levels for next trips (model 3-2).

Result

The framework is developed using random forest method and the results are compared to the observed values. For developing random forests R software is utilized. Initiating Random forest code in R needs some parameters to be defined including: number of trees in the forest, and number of explanatory variables that should be checked for splitting at each node. To make the forest development more efficient, especially when there are plenty of explanatory variables, the R code does not check all of the possible splits at each node to achieve the best split. Based on some researches it is suggested to check the square root of the total number of predictors' variables which have been selected randomly (4). Hence in this study this parameter is set as default which is the square root of total explanatory variables. For instance in the model 1 this parameter is set to 5 since we have 22 explanatory variables. Beriman (4) has proved that increasing number of trees in a random forest would not make it overfit and here this value is set to 1000 for all of the models.

Checking the goodness of fit can be examined through two dissimilar approaches. Since the proposed models are developed based on individuals, their accuracy in predicting correctly for each individual can be considered as the goodness of fit. On the other hand, the aggregate distribution for target variable and its difference from the observed distribution can be considered as another evaluation method. Obviously if the model is individually accurate it results in an aggregate distribution close to the observation, however the reverse is not always true. For a travel demand modelling it is vital to have good aggregate results, so we are not seeking maximizing individual accuracy. Consider trip generation model: according to observed independent variables we can classify peoples in a way that similar observations are in the same group. Nevertheless, still there is diversity in each group due to unobserved causes. For maximizing individual accuracy the model should associate the most observe value in each group to all the individuals, while a transport demand modeller tends to keep the diversity in his modelling and he might use the observed distribution in each group for predicting final result. In this way he deliberately has decreased the individual accuracy while he has improved the aggregate distribution.

In this research, both aggregate and disaggregate accuracies are presented. OOB error rate is the disaggregate criteria which is calculated based on that part of forest for developing which the data was excluded. Table 1 exhibits the OOB error rate for five developed random forests, based on which, the greatest error rate is for model 2-2 which has quite large number of target levels. Target levels represent the number of possible states for the target variable. Basically more number of levels results in more complicated modelling.

Table 1 – OOB error rate and variable features for models

Model	Number of target levels	Number of explanatory variables	OOB error rate (percent)
1	20	22	63.55
2-1	30	23	55.61
2-2	219	25	71.07
3-1	29	30	49.44
3-2	238	32	41.04

For the aggregate criteria, the marginal observed and predicted distributions for all of the independent variables are derived and the Kolmogorov-Smirnov (KS) test is utilized to check whether they are the same (10). Figure 2 display the observed and predicted marginal distributions and Table 2 presents the ks test's results for them. Based on this table for all the five marginal distribution we can accept that the observed and predicted distributions are the same.

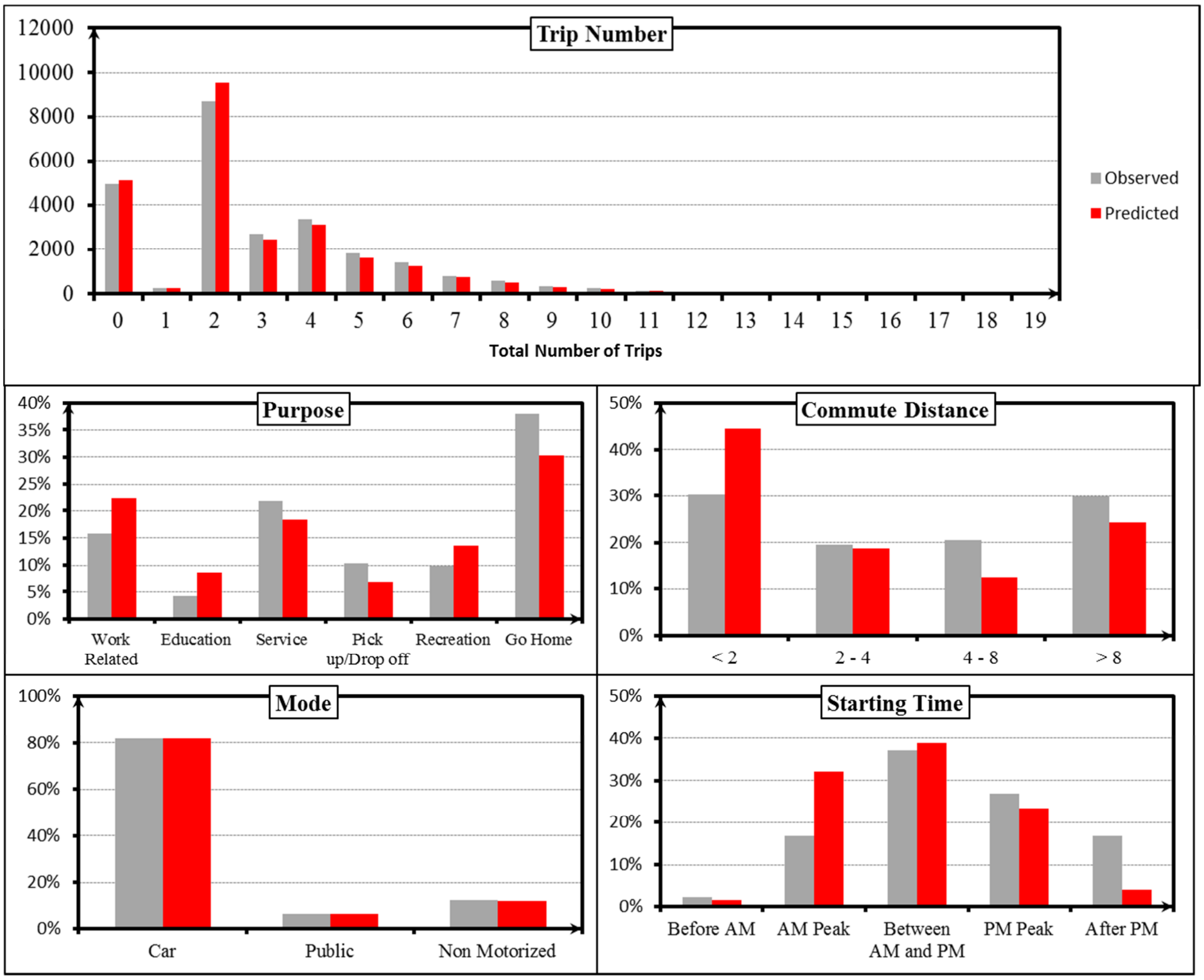

Figure 2 – Marginal observed and predicted distributions

Table 2 – Kolmogorov – Smirnov test's results for marginal distributions

Distribution	Number of categories	Maximum difference	Critical value
Total Number of Trips	20	0.042	0.304
Trip's Purpose	6	0.109	0.555
Trip's Starting Time	5	0.162	0.610
Trip's Mode	3	0.005	0.790
Commute Distance	4	0.143	0.680

Conclusion

This research proposes an innovative framework which is capable to predict individuals' trip diary in a quite straightforward approach. In this framework five separate models are developed that are simulated consecutively for completing people's daily travel pattern. The structure is consisted of three main steps: 1) Predicting total number of trips for individuals, 2) estimating attributes of the first trips, and 3) predicting attributes of the next trips. Each model uses the previous model's output and its result would be used as an input for the next one. Using the random forest data mining approach enhanced the accuracy and goodness-of-fit of the model by providing the possibility of incorporating a large set of explanatory variables and estimating unbiased estimates. Although the proposed framework does not necessarily replicate how individuals behave as much as an activity-based model does, complexity of decision making for several travel attributes are accounted for. In this study, destination attributes for each trip is addressed however the exact location of it is left for further research.

VISTA 2009 version of data is joined to SEIFA land use indexes to prepare a dataset for evaluating the framework. Kolmogorov-Smirnov test for marginal observed and predicted distributions as an aggregate capability, and OOB error ration as a disaggregate goodness of fit are checked to evaluate the results. The largest OOB error rate is about 70 percent in model 2-2 for which there are about 220 levels of target variable. According to ks results the argument that all the marginal predicted distribution are the same as their corresponding observed one can be accepted with the 95 percent confident level.

References

1 Kitamura R. (1988) An evaluation of activity-based travel analysis, *Transportation*, 15(1): 9-34

2 Axhausen, K., Garling, T.: Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. Transp. Rev. 12(4), 323–341 (1992)

3 Pendyala R., C. Chiu Y., P. Waddell, M. Hickman, K. Konduri and B. Sana, The design of an integrated model of urban continuum-location choices activity travel behavior and dynamic traffic patterns, 12th 5 WCTR, July 11-15, 2010 – Lisbon, Portugal

4 Breiman, L. (2001) Random Forests, *Machine Learning*, 45 (1): 5–32

5 Mohammadian, A., Zhang, Y. (2007) Investigating transferability of national household travel survey data, *Transportation Research Record,* 1993: 67–79

6 Rashidi T.H., Mohammadian A.(2011), Household travel attributes transferability analysis: application of a hierarchical rule based approach, *Transportation*, 38:697–714

7 DOT – Department of Transport (2009) Victorian Integrated Survey of Travel and Activity 2007, Melbourne: DOT

8 DOT – Department of Transport (2011) http://www.transport.vic.gov.au/vista, VISTA website Department of Transport, last accessed 10/22/2011

9 Transport NSW, Transport Data Centre (2010) 2008/09 Household Travel Survey Summary Report 2010 Release; REPORT 2010/01, JUNE 2010; Sydney: Transport NSW

10 Eadie, W.T., Drijard D., James F.E., Roos M. and Sadoulet B. (1971). Statistical Methods in Experimental Physics. Amsterdam: North-Holland, 269-271