

# The Application of Modal Split Using Revealed and Stated Preference Techniques: A study in Malang

Muhammad Nurjati Hidayat<sup>1</sup>

<sup>1</sup> Water Resources Engineering Department, Universitas Brawijaya, Malang, 65145, Indonesia

muhammadhidayat@ub.ac.id

Received 05-02-2018; revised 29-03-2018; accepted 08-04-2018

**Abstract.** In this study we investigate modal split and travel behaviour in Malang by conducting person trip survey in study area. The purpose is to understand respondents' travel behaviour and their preferences in selecting mode of transport. This is carried out to understand what are respondents feel regarding their perception on mode of transportation that available to them. The data being used are Revealed Preference (RP) and Stated Preference (SP) data. The first data based on present situation of respondents (including respondents' characteristics and daily travel information), while the second one is hypothetical scenario that has not available in present condition. These data then compared and analysed using Multinomial Logit Model (MNL).

Keywords: Travel behaviour, Revealed Preference, Stated Preference, Multinomial Logit Model

## 1. Introduction

In modelling travel demand, actual behaviour of user usually estimated using Revealed Preference (RP) data by employing discrete choice analysis (e.g. Ben-Akiva and Lerman [1]). However, RP data may have shortage in estimating individual choice due to the following reasons [2]:

- Preferences for non-existing services are not provided in RP data
- The set of alternatives considered by an individual may be ambiguous
- There are some errors in estimating service attributes
- There are similarities in attributes or lack of variability, or both

In order to lighten those drawbacks, a survey with hypothetical choice scenario and fully controlled alternatives need to be done. It called Stated Preference data that widely used by researchers for travel demand [3] and [4], and also in marketing research [5], as well.

This paper investigates the idea of using RP and SP data simultaneously because they have complementary characteristics. Unknown reliability of SP data is explicitly considered and its objective is to yield more reliable travel demand model on combining RP and SP data rather than analysing RP and SP data separately. The key features of this method are bias identification (the effect of new services that are not recognizable by individual in RP data), efficiency (all preferences on available data is jointly estimated), and bias correction [6].

In RP data, trade-offs among certain attribute cannot be estimated accurately. For example, the correlation between travel cost and travel time in RP data may resulting insignificant parameter

estimates for their coefficients. However, an SP survey designed based on little or low correlation between these attributes may presenting additional information on their trade-offs.

The aim of this paper is to perform the effectiveness of the combined RP/SP estimation method by an application to predict the mode choice in Malang (East Java) if new mode of transport is introduced.

## 2. Material and Methods

### 2.1. Study Area

This research located in Malang City (urban area) and Malang Regency (rural area) which is unified in Greater Malang territory (Batu City excluded in this research). Malang is the second biggest city in East Java having population of 820,243 residents, while Malang Regency having 2,459,982 residents. Total area is 110.06 and 3,534.86 km<sup>2</sup>, respectively. Government planning to introduce commuter train as a new public transportation in order to reduce the congestion level. Another reason is there is no public transport that connects directly from northern to southern part of Malang. By using the current transportation, it would take longer time rather than using private vehicle. On the ground of those reasons, the study of travel demand is needed.

### 2.2. Modelling Approach

Discrete choice model is a technique in which decision makers choose an alternative from choice set of available alternatives. It identifying pattern made by an individual facing available choice set. Discrete choice model postulates that the probability of individual choosing an alternative is based on their socioeconomics and level of attractiveness to the alternative. To represent the attractiveness of the alternative the utility function is constructed. Generally, utility derived from individual characteristics.

Utility is value indicator for an individual, which, generally, derived from the attributes of alternatives or sets of alternatives. The utility theory states that an individual will select an alternative in choice set that maximize his/her utility. The theory states that it contains function of attributes of alternative and individual characteristics[7]. In the newest study, the decision utility (mode choice) could be affected by characteristics the built environment (diversity, design and density) and travel modes (travel cost and time)[8].

The utility function,  $U$ , has a property that an alternative is chosen if its utility is greater than utility of other alternatives in individual's choice set. This function can be expressed as:

$$U_{in} \geq U_{jn}, \forall j \in C_n, i \neq j \quad (1)$$

$$U_{in} = V_{in} + \varepsilon_{in} \quad (2)$$

which could be interpreted as the utility of alternative  $i$  for individual  $n$  is greater or equal to alternative  $j$  in individual's choice set.

Utility function can be represented by two components: an observable or representative part  $V_{in}$  which is a function of the measured attribute, and a random term  $\varepsilon$  which reflect unobservable part of individual and also error made by modeler. Two individuals with same attribute,  $x$ , in observable part and having the same choice set probably select different choice, and some individual may not always choose the best alternative.

### 2.3. Model Specification

RP and SP model types are considered. The RP model reflecting behaviour of individual explained in utility function, whereas SP model is the result of SP response. Suppose mode choice model utility function for RP model with following equation:

$$U_{in}^{RP} = v_{in}^{RP} + \varepsilon_i^{RP} = \alpha' w_i^{RP} + \beta' x_i^{RP} + \varepsilon_i^{RP} \quad (3)$$

and the choice of decision maker given by:

$$d_{in}^{RP} = \begin{cases} 1, & \text{if } U_i = \max_{j=1, \dots, I} \{U_j\} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where  $i = 1, \dots, I$  is number of alternatives;  $w_i$  and  $x_i$  are vector of attributes of alternative;  $\alpha$  and  $\beta$  are parameter to be estimated; and  $\varepsilon_i$  is error term. Similar to utility function of RP model, the utility function of SP model is:

$$U_{in}^{SP} = v_{in}^{SP} + \varepsilon_i^{SP} = \beta' x_i^{SP} + \gamma' z_i^{SP} + \varepsilon_i^{SP} \quad (5)$$

and the choice of decision maker given by:

$$d_{in}^{SP} = \begin{cases} 1, & \text{if } U_s = \max_{j=1, \dots, I} \{U_j\} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

From the framework above we assume that SP response is the most preferable alternative chosen by respondent. Thus we have equation as follow:

$$U_{cn} \geq U_{an} \quad (7)$$

$$U_{tn} \geq U_{cn} \quad (8)$$

The subscripts  $c$ ,  $a$ , and  $t$  represent the current mode, alternative mode, and commuter train, respectively. The term  $x_i^{RP}$  and  $x_i^{SP}$  in both model implies the common variable in RP and SP model and term  $\gamma'z$  is specific to SP model that may contain SP biases. The level of random noise in RP and SP data is presented by the variance of the disturbance term  $\varepsilon$ . If the data source has different noise level, it can be shown as follow:

$$\text{Var}(\varepsilon_{in}^{RP}) = \mu^2 \text{Var}(\varepsilon_{in}^{SP}), \forall i, n \quad (9)$$

With the framework described above, the choice probability of alternative using multinomial logit model can be expressed as follow:

$$P_n(i) = \frac{\exp(v_{in})}{\sum_{j=1}^I \exp(v_{jn})} \quad (10)$$

#### 2.4. Combination of RP/SP Models

Since our concern is actual behaviour of respondents, prediction only using RP model. Hence, utility component being used is:

$$v_{in} = \hat{\beta}' x_{in} + \hat{\alpha}' w_{in} \quad (11)$$

Note that  $\hat{\beta}$  is calculated using both RP and SP data. In SP questions, hypothetical services are to be included for predicting demand, thus the term in SP model will be:

$$v_{in} = \hat{\beta}' x_{in} + \hat{\alpha}' w_{in} + \hat{\gamma}' \bar{z}_{in} \quad (12)$$

where  $\bar{z}_{in}$  represents hypothetical attributes related to the policy changes and  $\hat{\gamma}'$  is an estimation on parameter of  $\bar{z}_{in}$ . In equation above, RP and SP utility can be combined because new scale parameter ( $\mu$ ) is introduced with the purpose to adjust RP and SP scale parameters.

### 3. Result and Discussion

#### 3.1. Field survey

In this research we conducting two times survey, namely: preliminary survey and primary survey. The first survey conducted to ensure the population is being sampled, evaluate the questionnaire, and

the result is being evaluated. This survey also helps to avoid the problem of collecting large of useless or incorrect data in primary survey because of ineffective constructing questionnaire and improper sampling preparation. This survey targeting car, microbus, motorbike, and train users with 80 respondents. 20 passengers interviewed each of those modes, thus we understand their trip behaviour. Private vehicle (car and motorbike) users were interviewed in gas station, traffic light or when the user stops near the streets. Microbus users interviewed on-board and the rests was sampling people on train stations.

The second survey, primary survey, was conducted with small revision from preliminary survey. The sample being taken in this survey is 360 respondents for car, microbus, and motorbike. We eliminate train user because in preliminary survey there were only few respondents who use train as their mode for working purpose.

### 3.2. Result

The survey result is displayed in table 1. Number of respondent who travelled from Lawang or Kepanjen to city center is dominated by male (61%) compared to female (39%). Sequentially, the age of respondents are 25-45 years old (58%), 46-65 years old (22%), less than 25 years old (19%), and more than 65 years old (1%). The occupation of respondent is dominated by businessman (47%) and followed by entrepreneur (18%), civil servant (16%), student (10%), house wife and etcetera have same proportion, that are 4%, part timer has very small proportion (1%). Number of commute in a week conducted by respondents are 7 times a week (58%), 5 times (20%), more than 7 times (15%), and 3 times (7%). By taking notice to respondent who majority have job as businessman, probably there is correlation between occupation and number of commute in a week, that is 7 times a week.

Respondents' trip characteristics and level of service of transport facilities are shown as follow. The reason respondent not using public transportation dominated by long travel time (37%) followed by long waiting time (21%), uncomfortable (20%), having far access from home (11%), travel cost is expensive (11%), and etcetera (5%).

**Table 1.** The result of respondents' characteristics

| Item           | Range              | %  | Item                              | Range            | %  |
|----------------|--------------------|----|-----------------------------------|------------------|----|
| Sex            | Male               | 61 | Vehicle Ownership (%)             | Motorbike        | 70 |
|                | Female             | 39 |                                   | Car              | 30 |
| Age            | < 25               | 19 | Income (USD)                      | < 88             | 29 |
|                | 25 - 45            | 58 |                                   | 88 - 177         | 27 |
|                | 46 - 65            | 22 |                                   | >177             | 44 |
|                | 65 >               | 1  | People in Group                   | 1                | 67 |
| Occupation     | Civil Servant      | 16 |                                   | 3-Feb            | 23 |
|                | Entrepreneur       | 18 | >3                                | 10               |    |
|                | Businessman        | 47 | Reason not using public transport | Long travel time | 37 |
|                | Student            | 10 |                                   | Expensive        | 11 |
|                | Part Timer         | 1  |                                   | Uncomfortable    | 20 |
|                | House Wife         | 4  |                                   | Far access       | 11 |
|                | Others             | 4  | Long waiting time                 | 21               |    |
| Last Education | Elementary school  | 7  | Frequency per week                | ≤3               | 7  |
|                | Junior High School | 14 |                                   | 5                | 20 |
|                | Senior High School | 45 |                                   | 7                | 58 |
|                | University         | 27 |                                   | >7               | 15 |
|                | Post Graduate      | 5  |                                   |                  |    |
|                | Others             | 3  |                                   |                  |    |

**Table 2.** Respondents current and alternative mode and their income

| Current mode | Alternative mode |          |           |       | Income     |               |             | Total |
|--------------|------------------|----------|-----------|-------|------------|---------------|-------------|-------|
|              | Car              | Microbus | Motorbike | Total | Low income | Middle income | High income |       |
| Car          | 39               | 19       | 80        | 138   | 12         | 16            | 110         | 138   |
| Microbus     | 2                | 46       | 69        | 117   | 64         | 38            | 15          | 117   |
| Motorbike    | 10               | 52       | 83        | 145   | 39         | 51            | 55          | 145   |
| Total        | 51               | 117      | 232       | 400   | 115        | 105           | 180         | 400   |

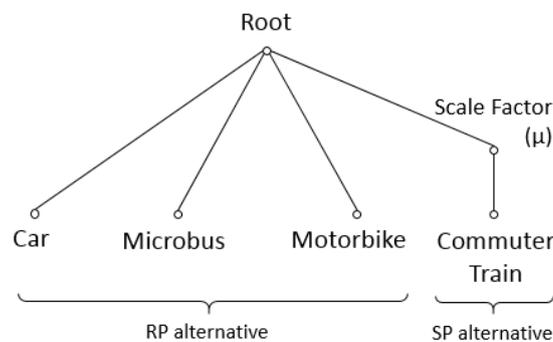
From Table 2 we can relate that respondents whom using car as their current mode prefers to use microbus and motorbike as their alternative mode are 19% and 80%, respectively. For microbus user, their alternative modes are car (2%) and motorbike (69%), while motorbike user prefers car (10%) and microbus (52%) as their alternative modes. Diagonally, we can see that there are respondents answered the alternative mode exactly the same as current mode, i.e. car respondent using car as their alternative, microbus respondent using microbus as their alternative and motorbike user using motorbike as their alternative. They are categorized as captive respondent, that is respondent who has no any alternative mode if current mode cannot be used.

Similar to the current mode of respondent over the alternative mode, income of respondents affecting modal choice preferences. From survey result, 80% of car respondent have high-income. Middle and high-income for motorbike respondent have almost identical numbers, that are 51 and 55 respondents, respectively. In contrary, microbus respondents have the opposite result of income. Majority of microbus user is a low-income respondent, which is 64 respondents (55%).

Afterwards, we could say that car user who uses motorbike as their alternative mode of transport is high-income respondent. Microbus user who categorized as captive respondent are middle-income. Only few microbus users who uses car categorized as high-income respondent. Motorbike respondent who uses microbus as their alternative mode mostly are middle-income respondent. High-income motorbike users mostly are captive respondent.

### 3.3. RP/SP Model Estimation

Three models were estimated: RP model, SP model and RP/SP model using Multinomial Logit (MNL) model, each of which was estimated by maximizing the log-likelihood function. In order to ease the explanation, artificial tree structure has been made. Given the way the data was collected, the structure in Figure 1 should represent these condition appropriately.



**Figure 1.** The artificial Tree Structure

**Table 3.** The result from estimation data

|                            | RP model       | SP model        | RPSP model      |
|----------------------------|----------------|-----------------|-----------------|
| Constant <sub>Car</sub>    | 2.155 (5.97)   |                 | 1.201 (12.85)   |
| Constant <sub>Bus</sub>    | 1.089 (1.59)   |                 | 0.953 (3.99)    |
| Constant <sub>Train</sub>  |                | 3.214 (7.75)    | 0.267 (10.59)   |
| Cost                       | 0.008 (2.04)   | -0.061 (-10.81) | -0.005 (-11.27) |
| Time                       | -2.927 (-2.68) | -3.843 (-7.05)  | -0.302 (-6.93)  |
| Age                        | 0.025 (1.33)   | -0.013 (-2.72)  | 0.0168 (2.65)   |
| Sex                        | -1.381 (-3.46) | 0.345 (3.22)    | -0.904 (-6.23)  |
| Frequency <sub>Train</sub> |                | 0.123 (10.88)   | 0.0098 (9.59)   |
| Income                     | -0.144 (-1.71) | 0.013 (0.96)    | -0.121 (-4.45)  |
| $\mu$                      |                |                 | 11.682 (10.38)  |
| $\rho^2$                   | 0.744          | 0.164           | 0.408           |
| N                          | 400            | 2000            | 2000            |

From RP data, a binary choice data set was created by regarding the current mode and the alternative mode as the first and the second options of chosen mode. The data set consist of 400 respondents, after eliminating incomplete and data error. In this data set, data collected only available in two alternatives, namely the current mode and the alternative mode. The current mode is the mode that respondent used every day for commuter trip, while the alternative mode is his/her alternative mode if the current mode is cannot be used. The current and alternative mode considered as the best choice or have the highest utility for respondent from the choice set.

The estimation of RP, SP and RP/SP model presented in Table 3. Not all models having correct signs due to some errors. Cost variable has sign that should not be positive. Logically, cost variable should have negative sign which means the higher the travel cost, the most likely respondent not using the current mode as their transport mode. The explanation of positive sign in cost variable could be as follow:

1. There is a correlation between distance and travel fare. The longer the distance of the travel, the higher travel fare should be paid by respondent
2. The current mode has higher cost than the alternative mode. The expensive current mode is the product of the number of respondent who have travel expenses using the current mode higher than the alternative mode divided by number of non-captive respondents and multiplied by hundred to obtain the results in percentage. Captive respondent excluded in the calculation because there is no significant difference in travel cost.
3. There were mistakes in translating questionnaire form from English version into Indonesian version before distribute it to surveyor.

As for SP model, stated preference or stated intention of using the new commuter train were employed to create binary choice data. If the respondents have willingness or intention to switch mode using commuter train, he or she is considered to have chosen the commuter train over the current mode he/she currently used. Otherwise, the respondent is considered to choosing the currently mode they used over the commuter train. Thus, a binary choice can be created. In the data, each respondent facing five different level of services of new commuter train. With number of observation of 2000, respondent will accept and have intention to use new commuter train if the level of service match his/her condition.

Commuter train constant is introduced in the model so that it may capture the attributes of new commuter train that are not include in the model and response bias toward the new commuter train. The commuter train constant has a significant positive coefficient, probably reflect the overstated use of the commuter train.

The last model is combined RP/SP model. The framework is shown in equation (3) to (9). Due to the difference in variance of RP and SP data, variable  $\mu$  is employed. The value of scale parameter  $\mu$  is expected to be less than one. If scale parameter  $\mu$  has value less than one, it scales down the explanatory variables in the stated preference model because SP model has more random noise than RP model. However, if the  $\mu$  value greater than one, the RP model has more random noise.

The estimation results of RP/SP model are shown in the third column of table 3. It shows that train frequency has positive coefficient. The results also show that  $\mu$  has estimated value greater than 1. This probably because of the errors obtained from RP data explained previously. The value more than 1 scales up the SP model. The estimation of RP/SP model is almost similar to RP model which indicate that the joint estimation successfully replicates the RP model, except in variable cost.

#### 4. Conclusions

The combined estimation of discrete choice models from RP and SP data was presented. The strategy combining both types of data can benefit with explicit consideration of their merit and demerit. The case study is modal split model under hypothetical scenario, namely introducing commuter train as new alternative. In estimating RP and SP data simultaneously to estimate the mode choice model, alternative specific constant was estimated separately. In modelling using MNL model, the most significant model from RP model, SP model and combined RP/SP model is RP/SP model than other models alone. Combining RP data with SP data increase the accuracy of parameter estimates in the model. In our result shows that RP model contain more random noise than SP model. To perform this model in real condition, more data need to be collected, such as number of passenger for each mode and OD table, thus the probability for each mode can be calculated. From academic point of view, this research need more advanced model such as Nested Logit (NL), Cross Nested Logit (CNL), Generalized Nested Logit (GNL) to resulting higher quality in estimation and research due to the estimated modes are able to categorized in private vehicle and public transport.

#### References

- [1] M. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, 1985.
- [2] T. Morikawa, M. Ben-Akiva, and K. Yamada, "Forecasting intercity rail ridership using revealed preference and stated preference data," *Transp. Res. Rec.*, vol. 1328, pp. 30–35, 1991.
- [3] J. J. Louviere *et al.*, "Laboratory-Simulation Versus Revealed-Preference," *Transp. Res. Rec.* 794, pp. 42–51, 1981.
- [4] D. a. Hensher, P. O. Barnhard, and T. P. Truong, "The role of stated preference methods in studies of travel choice," *J. Transp. Econ. policy*, vol. 22, no. 1, pp. 45–58, 1988.
- [5] P. Cattin and D. R. Wittink, "Commercial Use of Conjoint Analysis: A Survey," *Source J. Mark.*, vol. 46, no. 3, pp. 44–53, 1982.
- [6] M. Ben-Akiva *et al.*, "Combining revealed and stated preferences data," *Mark. Lett.*, vol. 5, no. 4, pp. 335–349, 1994.
- [7] T. A. Domencich and D. McFadden, *Urban Travel Demand: A behavioral Analysis*. Amsterdam and Oxford, 1975.
- [8] J. De Vos, P. L. Mokhtarian, T. Schwanen, V. Van Acker, and F. Witlox, "Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility," *Transportation (Amst.)*, vol. 43, no. 5, pp. 771–796, 2016.