Modal choice models estimation using mixed Revealed and Stated Preferences data

R. Guzzo & G. Mazzulla  
*Dipartimento di Pianificazione Territoriale, University of Calabria, Italy*

**Abstract**

Discrete choice demand models estimation can be made using information on travel behavior obtained by Revealed and Stated Preferences surveys.

  Revealed Preferences techniques, traditionally utilized, are relative to the actual users travel behavior in a real context.

  Stated Preferences techniques are methodologies based on statements made by interviewees about their preferences in different choice contexts, real, hypothetical or experimental. Therefore, an important innovation is introduced: the possibility to consider choice alternatives not available at the time of the surveys.

  These techniques have been adopted for demand models calibration to predict the choices made by users and their preferences variations while the choice context changes. Particularly, some Multinomial Logit mode choice models have been specified and calibrated.

  The calibrated models have been distinguished in: RP models, based on the choices made by users exclusively in the real context; SP models, based on the choices stated in the hypothetical contexts; and joint RP/SP models, using RP and SP data on the same sample.

  The study has confirmed the utility of SP techniques for users travel behavior analysis in a hypothetical choice contexts; moreover, it has confirmed, as attended, that RP and SP conjoint analysis improves parameters estimation in discrete choice models.

*Keywords: discrete choice models estimation, Revealed Preferences, Stated Preferences, joint calibration.*
1 Introduction

Traditionally, the mobility surveys are made with the Revealed Preferences method (RP), relative to the actual users travel behavior in a real context. Since the early 1970s, some marketing researchers devised new surveys methodologies, known in literature as Stated Preferences techniques (SP); these techniques were immediately of great interest also for transport researchers; the first applications in this field began in the early eighties.

Stated Preferences techniques are methodologies based on the statements made by interviewees about their preferences in different choice contexts, real, hypothetical or experimental. Therefore, an important innovation is introduced: the possibility to consider choice alternatives unavailable at the time of the surveys [9].

These techniques have been adopted for the demand models calibration to predict the choices made by users and their preferences variations while the choice context changes.

2 SP data analysis

SP surveys methodologies involve the definition of [5]:
1. choice alternatives;
2. attributes (or factors) considered for each alternative;
3. levels of variation of each attribute;
4. choice contexts (scenarios) proposed to the decision-maker;
5. type of preference asked;
6. modality of interviews management.

The number of possible scenarios depends on the combinations of numerosity among the number of alternatives, the number of attributes and the number of levels of each attribute. The type of preference can be different: choice, ranking and rating.

In the first case, the alternative chosen in that context must be indicated; in the second case, the available options must be ordered according to the degree of preference; in the third case, a preference measure must be assigned to each alternative, according to a predefined semantic scale.

Whichever type of preference used in SP survey design, the analyst’s aim is to establish the relative effect of each attribute on the overall utility that individuals associate to each option. For this aim, the most commonly applied analytical approaches are the following [9]:
- the approach of “probabilistic” discrete choice models, based on choice probability of each option regarding all other available; the most common form of such models is the “logit” function;
- the approach of regression models, adapted to analyse ranking and rating data;
the approach of analysis of variance, adapted to analyse ranking data, which allows to estimate users’ preferences structures; this approach introduces a low statistical reliability, that’s why it is not used much.

The SP data applications, related to stated choices in a particular hypothetical context, have assumed a growing importance in the last few decades. Some authors have proposed methodologies to use this kind of data and models derive by them [7, 9]; however, many authors assert that a direct application of these models to forecast the choices made by the users isn’t very appropriated [3, 6]; some authors have proposed joint calibration models using RP and SP data [1, 4].

The collected data through SP surveys can be used for demand models calibration relative to choice dimensions proposed to decision makers. Substantially, the estimation methods that can be adopted are those used for demand models calibration using RP surveys, but they differ for the type of preference used on SP experiment.

The choice data can be analysed as RP data; the only difference could consist in considering some hypothetical choice alternatives regarding those “revealed” in RP surveys. Also in ranking data the information can be analysed in traditional way, considering alternatives two by two, and dealing the answers as if the “choice” was the alternative classified as better between two successive answers. Finally, in rating data, it is possible to assume as potentially chosen the alternative with higher value in the semantic scale, or to consider the “rate” associate to each alternative as indicative of utility.

In the case of joint calibration using RP and SP data, usually it is applied the scaling estimation methodology; this methodology allows to consider the variability among different types of data used jointly in a statistic analysis. For discrete choice models, a possible specification of the perceived utilities by users in real (RP) and hypothetical (SP) context is the following [5, 8]:

\[ U_{i}^{RP} = \beta \cdot X_{i}^{RP} + \alpha \cdot Y_{i}^{RP} + \epsilon_{i}^{RP} \]  
\[ \theta \cdot U_{i}^{SP} = \theta(\beta \cdot X_{i}^{SP} + \gamma \cdot Z_{i}^{SP} + \epsilon_{i}^{SP}) \]

where:

- \( U_{i}^{RP}, U_{i}^{SP} \) are the perceived utilities by the user \( i \) in the RP and SP context, respectively;
- \( X_{i}^{RP}, X_{i}^{SP} \) are the vectors of the observed values of the common variables to the RP and SP data, respectively;
- \( Y_{i}^{RP}, Z_{i}^{SP} \) are the vectors of the observed values of the specific variables to the RP or SP data, respectively;
- \( \alpha, \beta, \gamma \) are the vectors of the parameters to be estimated, in both RP and SP data sets;
- \( \epsilon_{i}^{RP}, \epsilon_{i}^{SP} \) are the vectors of the random residuals; they represent the unexplained variability of the utility function, in both RP and SP data sets;
\( \theta \) is the scale factor, so that: 
\[
\theta^2 = \frac{\text{var}(\epsilon_i^{RP})}{\text{var}(\epsilon_i^{SP})}
\]

When the random residuals \( \epsilon_i^{RP}, \epsilon_i^{SP} \) are distributed according to the Gumbel function, with zero mean but a different variance among RP and SP data, the choice probabilities \( P_i^{RP} \) and \( P_i^{SP} \) of the observation \( i \) (interviewed \( i \), in choice data or scenario in ranking or rating data) are:

\[
P_i^{RP} = \frac{e^{(\beta X_i^{RP} + \alpha_i^{RP})}}{\sum_j e^{(\beta X_j^{RP} + \alpha_j^{RP})}} \tag{3}
\]

\[
P_i^{SP} = \frac{e^{\theta(\beta X_i^{SP} + \gamma_i^{SP})}}{\sum_j e^{\theta(\beta X_j^{SP} + \gamma_j^{SP})}}
\]

The joint estimation of the parameters \( \beta, \alpha, \gamma, \theta \) can be obtained maximizing the Likelihood function of the joined sample, with the hypothesis that the two samples are independent:

\[
L(\beta, \alpha, \gamma, \theta) = \left( \prod_{n=1}^N \left( \prod_{A \in \Lambda(q)} P_{iq}^{RP} \right) \right) \left( \prod_{n=1}^N \left( \prod_{A \in \Lambda(q)} P_{iq}^{SP} \right) \right)
\]

\[
(4)
\]

The function is non-linear because \( \theta \) multiplies not only the attributes, but also the parameters of the SP utility function. Consequently, the estimation procedure of the scale factor is simply an operational research problem, that can be resolved through a specific software able to manage non-linear Likelihood functions directly, or through techniques that allow the use of non-specific software and developed for discrete choices analysis. In the last case, in literature three methods are brought substantially: the first one, proposed by Ben-Akiva and Morikawa [1], based on a sequential estimation procedure; the second one, proposed by Bradley and Daly [2], based on the simultaneous estimation of the RP/SP model parameters and the parameter \( \theta \); the third one, proposed by Postorino and Pirrello [10], based on an iterative estimation procedure.

The authors Ben-Akiva and Morikawa have proposed a sequential estimation method; the algorithm consists to estimate a SP model initially, with the aim of estimating the parameters \( \theta \beta \) and \( \theta \gamma \), corrected according to the factor \( \theta \). Then these parameters are used for defining a new variable, to be included in the utility of the choice alternatives of the RP model:

\[
U_i^{RP} = \lambda V_i^{RP} + \alpha Y_i^{RP} + \epsilon_i^{RP}
\]

where \( \lambda = 1/\theta \).
Using an usual estimation procedure, the parameter $\lambda$ can be estimated and, consequently, the scale factor can be calculated.

Multiplying the vectors of the attributes $X^{SP}$ and $Z^{SP}$ for the estimated parameter $\theta$, a modified SP data set is obtained. Combining the RP and SP data, the joint model can be estimated.

The authors Bradley and Daly propose a procedure that consists of constructing an artificial tree, in which there are two set of alternatives: a first set is constituted by RP alternatives, a second one is constituted by SP alternatives. RP alternatives are placed just below the root of the tree, while SP alternatives are each placed in a single-alternative nest. For an RP observation, SP alternatives are unavailable set for the user, and the choice is modelled as in a standard Logit model. For an SP observation, RP alternatives are unavailable set for the user, nevertheless the choice is modelled as in a nested Logit model.

For SP observations, the utility functions of each composite alternative are calculated as:

$$V^{COMP} = \theta \log \sum e^{v^{SP}} = \theta \log \sum e^{(\beta X^{SP} + \gamma Z^{SP})}$$

where the sum is taken over all of the alternatives in the "nest" corresponding to the composite alternative; particularly, each "nest" in this specification represents a single-alternative, and therefore:

$$V^{COMP} = \theta V^{SP} = \theta \beta X^{SP} + \theta \gamma Z^{SP}$$

The model structure assures that the value of $\theta$ is the same for each of the dummy alternatives.

The scale factor calculation can also be obtained with an iterative method proposed by Postorino and Pirrello. This method consists of fixing an initial value of the scale factor $\theta$, and to produce new vectors of the parameters $X_i^{SP} = \theta X^{SP}$ and $Z_i^{SP} = \theta Z^{SP}$; this allows to calculate the function $\log L^{(RP+SP)}(\beta, \alpha, \gamma, \theta)$ and to estimate the initial unknown parameters (model coefficients and scale factor). The log-Likelihood objective function is so valued iteratively in different points, each of which corresponds to a value of the parameter $\theta$, until it is defined the value that optimizes the same function; when the optimization value is obtained, the unknown parameters are correctly estimated.

3 Experimental survey in the university campus of Cosenza

It has been realized an experimental survey in the campus site of the University of Calabria, situated in the urban area of Cosenza (Italy). The campus is attended by 28,000 students and 2,000 members of staff approximately (March 2003). Currently the University is served by a bus service, which doesn’t resolve in a suitable way the students’ mobility demand; where possible, they prefer to use
the individual transport, producing congestion both on the access and on the internal campus road networks. The surveys, realized in the spring of 2003, have interested a sample of 281 Engineering Faculty students, on a total of 6,600 students approximately. The surveys have been planned to collect, on the same users’ sample, RP and SP data. RP surveys have been done to collect some users’ socio-economic characteristics and information regarding transport mode used to reach the university campus.

The SP experiment has been made, instead, to estimate a strategy aimed to increase the use of collective transport that connects the urban area with the campus; the strategy consists in conjugating policies to improve collective transport services with management demand policies, i.e. car-park pricing. In total, the users have made 341 tours with at least one stop in the university area; 221 users have made one tour in a day alone; 60 users have made two tours in a day. The access trip is realized primarily with an individual transport mode (74.2%), and only 25.8% of decision makers uses the collective transport.

In SP experiment, the users have expressed their degree of preference (in according to a semantic scale from 1 to 5) on 7 hypothetical choice scenarios; in each scenario both car and bus alternatives are present. The car alternative is characterized by the parking cost attribute, that varies by actual level (free) to the intermediate and elevated level; the bus alternative is characterized by frequency (low and high) and travel time attributes (equal or reduced regarding to actual time).

The results show that the users of the transport system are inclined to accept the policies of mobility government proposed when the use of car is strongly discouraged, through car-park pricing policies. In fact, improvement of collective transport services involves a meaningful variation in the modal split choice when an incisive additional cost is imposed at individual use of car.

4 Calibration of models for the mode choice with mixed RP and SP data

In this paper, some MNL mode choice models have been specified and calibrated [5], on the base of the users’ choices about their access trip to the university area.

The calibrated models have been distinguished in RP models, based on the choices made by users exclusively in the real context, and in SP models, based on the choices stated in the hypothetical contexts, characterizing as “choice” the alternative to which the users associate, in rating experiment, a greater degree of preference. In order to improve the estimation parameters obtained by previous elaborations, joint RP/SP models have been calibrated. The results of estimations are reported in tables 1 and 2.

The first specification proposed (table 1) has a simple structure, in which only two modal alternatives have been defined (car and bus); each alternative is characterized by the usual attributes of level of service (times and costs of travel, bus frequency). The time attribute, in minutes, represents the total time on the access trip for the car and bus alternatives. The cost attribute, in Euros,
represents, for the car alternative, the kilometric cost valued for the access trip added to the possible cost for the car-parking, and the total ticket cost for the bus alternative. The bus frequency has been expressed as a dichotomous variable of value equal to one for elevated frequency and zero otherwise.

Table 1: Model calibration results (I specification).

<table>
<thead>
<tr>
<th>Alternative</th>
<th>Variable</th>
<th>Parameters</th>
<th>Value estimated</th>
<th>t-Student</th>
<th>Value estimated</th>
<th>t-Student</th>
<th>Value estimated</th>
<th>t-Student</th>
<th>Value estimated</th>
<th>t-Student</th>
<th>Value estimated</th>
<th>t-Student</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RP Model</td>
<td>SP Model</td>
<td>RP/SP Model (a)</td>
<td>RP/SP Model (b)</td>
<td>RP/SP Model (c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAR</td>
<td>TTime</td>
<td>β₁</td>
<td>-0.0287</td>
<td>-2.309</td>
<td>-0.0212</td>
<td>-4.625</td>
<td>-0.0118</td>
<td>-4.993</td>
<td>-0.0120</td>
<td>-4.998</td>
<td>-0.0124</td>
<td>-2.013</td>
</tr>
<tr>
<td></td>
<td>TCost</td>
<td>β₂</td>
<td>-0.4240</td>
<td>-1.555</td>
<td>-1.0520</td>
<td>-11.110</td>
<td>-0.5494</td>
<td>-11.240</td>
<td>-0.5572</td>
<td>-11.230</td>
<td>-0.5765</td>
<td>-2.284</td>
</tr>
<tr>
<td>BUS</td>
<td>TTime</td>
<td>β₁</td>
<td>-0.0287</td>
<td>-2.309</td>
<td>-0.0212</td>
<td>-4.625</td>
<td>-0.0118</td>
<td>-4.993</td>
<td>-0.0120</td>
<td>-4.998</td>
<td>-0.0124</td>
<td>-2.013</td>
</tr>
<tr>
<td></td>
<td>TCost</td>
<td>β₂</td>
<td>-0.4240</td>
<td>-1.555</td>
<td>-1.0520</td>
<td>-11.110</td>
<td>-0.5494</td>
<td>-11.240</td>
<td>-0.5572</td>
<td>-11.230</td>
<td>-0.5765</td>
<td>-2.284</td>
</tr>
<tr>
<td>BUS-RP</td>
<td>α₁</td>
<td></td>
<td>-0.6357</td>
<td>-2.434</td>
<td>-0.9632</td>
<td>-6.509</td>
<td>-0.9610</td>
<td>-6.484</td>
<td>-0.9569</td>
<td>-5.957</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FreqBus</td>
<td>γ₁</td>
<td>-1.1600</td>
<td>10.490</td>
<td>0.6078</td>
<td>10.460</td>
<td>0.6173</td>
<td>10.460</td>
<td>0.6384</td>
<td>2.200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUS-SP</td>
<td>γ₂</td>
<td></td>
<td>-0.4624</td>
<td>-3.906</td>
<td>-0.2316</td>
<td>-3.756</td>
<td>-0.2343</td>
<td>-3.743</td>
<td>-0.2411</td>
<td>-2.014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V.O.T. (€/h)</td>
<td></td>
<td></td>
<td>4.07</td>
<td>1.21</td>
<td>1.29</td>
<td>1.29</td>
<td>1.29</td>
<td>1.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>θ</td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td>1.90</td>
<td></td>
<td>1.87</td>
<td></td>
<td>1.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogL(β)</td>
<td></td>
<td></td>
<td>-149.56</td>
<td></td>
<td>-1152.60</td>
<td></td>
<td>-1152.72</td>
<td></td>
<td>-1152.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogL(θ)</td>
<td></td>
<td></td>
<td>-189.23</td>
<td></td>
<td>-1165.57</td>
<td></td>
<td>-1355.10</td>
<td></td>
<td>-1355.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td></td>
<td></td>
<td>79.3 (χ²=7.81)</td>
<td>328.4 (χ²=9.48)</td>
<td>405.0 (χ²=11.07)</td>
<td>404.8 (χ²=11.07)</td>
<td>405.5 (χ²=11.07)</td>
<td>0.1938</td>
<td>0.1374</td>
<td>0.1494</td>
<td>0.1457</td>
<td>0.1496</td>
</tr>
</tbody>
</table>

The proposed formulations have concurred to easily verify the estimation procedure of the scale factor between two data sets, in relation to various procedures proposed in literature; in the first case (model a) an iterative estimation method proposed by Postorino and Pirrello has been applied [10]; in the second one (model b) a sequential estimation method proposed by Ben-Akiva and Morikawa has been applied [1]; in the third case (model c) a simultaneous estimation method proposed by Bradley and Daly has been applied [2].

The results show that all parameters have a correct sign and have a value statistically different from zero, to a good level of significance. As expected, the cost attribute coefficients have a negative sign, while the bus frequency is the only parameter that assumes positive sign.

All models verify the statistical tests on the goodness-of-fit. The statistics assume the maximum value in the RP model (equal to 0.1938), confirming the greater predictive capability of the model, which simulates the choices made in a real context; the joint model, instead, allows to improve the predictive capability to the SP model. In the proposed models, the estimation of the monetary value of time (VOT) is different; in the RP model the VOT is equal to 4.07 €/h, indicating that the user is disposed to pay such fee in order to save an hour of time.
However, considering the interviewed sample is composed of students, the fee could seem excessive; a more realistic estimation, therefore, could be that estimated by joint RP/SP model (equal to 1.29 €/h).

The results obtained from the different procedures of the scale factor converge on a factor equal to 1.9. The parameters estimated by different scale factor estimation procedures (RP/SP models) are similar even if the simultaneous estimation procedure leads to statistically non significant parameters. Moreover, the scale factors obtained converge to approximately equal values. The scale factor represents the relationship between RP and SP error variances; therefore, a value higher than one indicates a greater variability of the RP model due, probably, to omission of some attributes in the utility function. Furthermore, since scale factor deals with the correction of the parameter estimation in the SP model utility function, the obtained value indicates an over-estimation of these parameters and the necessity “to readjust” their value reducing it of a factor equal to 1.9.

In a successive phase, more complex model formulations have been proposed, with a variable that indicates the user’s income (variable from 1, low income, to 5, high income), a variable that indicates the user’s gender (variable dichotomous equal to one for female gender and equal to zero otherwise), and a variable indicative of the student’s “out-side” condition, which assumes value equal to one when the student is resident in a place distant from the university campus and zero otherwise (table 2).

Table 2: Model calibration results (II specification).

<table>
<thead>
<tr>
<th>Alternative</th>
<th>Variable</th>
<th>RP Model</th>
<th>SP Model</th>
<th>RP/SP Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameters</td>
<td>Value estimated</td>
<td>t-Student</td>
<td>Value estimated</td>
</tr>
<tr>
<td>CAR</td>
<td>TTime</td>
<td>$\beta_1$</td>
<td>-0.0264</td>
<td>-2.025</td>
</tr>
<tr>
<td></td>
<td>TCost</td>
<td>$\beta_2$</td>
<td>-0.6346</td>
<td>-2.080</td>
</tr>
<tr>
<td></td>
<td>Income</td>
<td>$\beta_3$</td>
<td>0.6486</td>
<td>5.745</td>
</tr>
<tr>
<td>BUS</td>
<td>TTime</td>
<td>$\beta_1$</td>
<td>-0.0264</td>
<td>-2.025</td>
</tr>
<tr>
<td></td>
<td>TCost</td>
<td>$\beta_2$</td>
<td>-0.6346</td>
<td>-2.080</td>
</tr>
<tr>
<td></td>
<td>Sex</td>
<td>$\beta_4$</td>
<td>0.9216</td>
<td>2.275</td>
</tr>
<tr>
<td></td>
<td>OutSide</td>
<td>$\beta_5$</td>
<td>1.894</td>
<td>5.812</td>
</tr>
<tr>
<td></td>
<td>FreqBus</td>
<td>$\gamma_1$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>V.O.T. (€/h)</td>
<td></td>
<td>2.50</td>
<td>0.79</td>
<td>0.92</td>
</tr>
<tr>
<td>$\theta$</td>
<td></td>
<td>-</td>
<td>-</td>
<td>0.72</td>
</tr>
<tr>
<td>LogL($\beta$)</td>
<td></td>
<td>-124.06</td>
<td>-970.79</td>
<td>-1109.79</td>
</tr>
<tr>
<td>LogL($\theta$)</td>
<td></td>
<td>-189.23</td>
<td>-1165.87</td>
<td>-1355.10</td>
</tr>
<tr>
<td>LR</td>
<td></td>
<td>130.3 ($\chi^2 = 11.07$)</td>
<td>390.2 ($\chi^2 = 12.59$)</td>
<td>490.6 ($\chi^2 = 12.59$)</td>
</tr>
<tr>
<td>$\rho^2$</td>
<td></td>
<td>0.3180</td>
<td>0.1622</td>
<td>0.1766</td>
</tr>
<tr>
<td>%RIGHT</td>
<td></td>
<td>78.55% (216/275)</td>
<td>69.9% (1199/1715)</td>
<td>70.65% (1401/1983)</td>
</tr>
</tbody>
</table>
In this case, all parameters have a correct sign and have a value statistically different from zero, except the “sex” variable. Particularly, the "income", "sex" and "out-side" parameters have assumed a positive sign, as attended for the defined value of variables.

In this second specification, the socio-economic variables in the utility function have made the alternative specific attribute (ASA) not statistically significant which, therefore, has been excluded by these models.

All models verify statistical tests on goodness-of-fit; joint RP/SP model, similar to previous results, improves the SP model estimation. The proposed specifications, moreover, improve the models forecasting capability, and in particular of the RP/SP model. In this second specification, the scale factor assumes a value inferior to one, indicating a better performance of the RP model and an under-estimation of the estimated parameters in the SP model.

5 Conclusions

The study has confirmed the utility of SP techniques for users travel behavior analysis in hypothetical choice contexts; it has confirmed, moreover, than RP and SP conjoint analysis improves the estimation of parameters in discrete choice models. The proposed models allow the forecast of the users’ preference variations while the level of the attributes of the proposed alternatives changes. In the research development the proposed models transferability is inquired, so that models can be used for forecasting in different contexts.

References


