Calibrating residential location choice model: SP and RP data

Luigi Biggiero - Francesca Pagliara

DEPARTMENT OF TRANSPORTATION ENGINEERING
UNIVERSITY OF NAPLES "FEDERICO II"

The objective of this paper is to explore the use of stated and revealed preference techniques for the calibration of residential location choice sub-model, being part of a larger land use and transport model. The aim is to observe attitudes of people and therefore obtaining information to guide planning activity.

Two examples of calibrating residential location choice model are described. In the first, a stated preference experiment was conducted in Calgary. In each observation each respondent was presented with a set of hypothetical residential location alternatives and asked to indicate the order of preference for these alternatives. In the second, a revealed preference survey was employed in the city of Naples where each interviewee had stated his/her zone of residence, of work.

1 Introduction

It has long been argued that transportation system, through its effects on accessibility, has various impacts on the attractiveness of locations as sites for activities.

More specifically, it has been asserted that the relative travel times and the ease of access provided by the roadway and the public transport systems serving an area influence the relative degrees of attractiveness individuals associate with different residential locations in the area (Wegner and Furst, 1999).

There is an extensive literature on the study of residential location choice behaviour in various urban areas. The content of this literature is especially in terms of the factors found to have influence. A wide variety of dwelling unit attributes location attributes and households characteristics have been shown to be factors influencing this behaviour. Most studies have found that money cost, dwelling unit size and proximity to activities have major influences. Similarly, household size, life cycle and income have often been identified as important characteristics. The nature of the observations of preference is not deeply treated in literature.

The objective of this paper is to explore the use of stated preference and revealed techniques in calibrating residential location choice models. Section 2 describes the procedure of calibrating random utility models; section 3 describes the use of SP and RP in calibrating residential location choice model. Section 4 and 5 propose two examples of calibrating such models with SP and RP data respectively. Section 6 offers conclusions that arise from the work.

2 Model calibration

Random utility models can be seen as mathematical relationships expressing the probability $p^i[j](X,\beta,\theta)$ that individual i chooses alternative j as a function of the vector X of attributes for all the available alternatives and of the vectors of parameters relative to the systematic utility, β, and to the joint probability function of the random residuals, θ (Cascetta, 2001). Choice probabilities depend on X and θ through systematic utility functions, usually specified as linear combinations of the attributes X (or their transformations) with coefficients given by the parameters β:

$$V_j(X_j^i) = \Sigma_z \beta_z X_{zj}^i = \beta^T X_j^i \quad (1)$$

Structural parameters θ include all parameters related to the probability distribution function of random residuals.

Calibrating the model requires the estimation of the vectors β and θ from the choices made by a sample of users.

The Maximum Likelihood Method is the method most widely used for estimating model parameters. In Maximum Likelihood estimation the values of the unknown parameters are obtained by maximising the probability of observing the choices made by a sample of users. The probability of observing these choices, i.e. the likelihood of the sample, depends (in addition to the choice model adopted) on the sampling strategy adopted. The case of simple sampling will be considered in the following.

In the simplest case of *simple random sampling* of n users, the observations are statistically independent and the probability of observed choices is the product of the probabilities that each user i chooses $j(i)$, i.e. the alternative actually chosen by him/her. The probabilities $p^i[j(i)](X^i; \beta, \theta)$ are computed by the model and therefore depend on the coefficients vectors. Thus, the probability L of observing the whole sample is a function of the unknown parameters:

$$L(\beta,\theta) = \Pi_{i=1,\dots,n}\, p^i[j(i)](X^i; \beta, \theta) \quad (2)$$

The Maximum Likelihood estimate $[\beta,\theta]_{ML}$ of the vectors of parameters β and θ is obtained by maximising (2) or, more conveniently, its natural logarithm (log-likelihood function):

$$[\beta,\theta]_{ML} = argmax\ lnL(\beta,\theta) = argmax\ \Sigma_{i=1,\dots,n} ln\, p^i[j(i)](X^i, \beta, \theta) \quad (3)$$

Once a demand model has been specified and calibrated, it must be validated. In this phase the reasonableness and the significance of estimated coefficients are verified, as well as the model's capability to reproduce the choices made by a sample of users. In addition, the assumptions underlying the functional form assumed by the model are tested. All of these

activities can be completed with appropriate tests of hypotheses for a sample of users.

Informal tests on coefficients. These tests are based on the expectations on the signs of the coefficients calibrated and on their reciprocal relationships.

Wrong signs of the coefficients are likely indicators of errors in the attribute database in survey results, or of the model mis-specification.

Formal tests on coefficients.

t-student tests on particular coefficients

These tests check the null hypothesis that a coefficient β_k is equal to zero and the estimate β_k^{ML} differs from zero for sampling errors ($H_0 : \beta_k = 0$) through the statistic:

$$t = \frac{\beta_k^{ML}}{\mathrm{Var}[\beta_k^{ML}]^{1/2}} \qquad (4)$$

Alternatively, the t-student statistic can be used to test that two coefficients β_k and β_j are equal ($H_0 : \beta_k = \beta_j$):

$$t = \frac{\beta_k^{ML} - \beta_j^{ML}}{(\mathrm{Var}\,[\beta_k^{ML}] + \mathrm{Var}\,[\beta_j^{ML}] - 2\,\mathrm{cov}\,[\beta_j^{ML}\,\beta_k^{ML}])^{1/2}} \qquad (5)$$

In both cases, under the null hypothesis the statistic t is distributed according to a t-student variable with a number of degrees of freedom equal to the size of the sample minus the number of coefficients estimated. Given the typical sample size it is usually assumed that the t statistic is distributed as a normal standard variable, $N(0,1)$, which is the limit distribution of the t-student variable. It is well known that the null hypothesis is rejected with a probability (of making α Type I error (e.g. rejecting a true assumption) if the value of the t statistic is external to the extremes of the interval ($z_{\alpha/2}$, $z_{1-\alpha/2}$) which for α=0.05 are equal to $\pm$ 1.96.

Statistics and tests on goodness of fit. The model's capability to reproduce the choices made by a sample of users can be measured by using the rho-square statistic:

$$\rho^2 = 1 - \frac{lnL\,(\beta^{ML})}{lnL\,(0)} \qquad (6)$$

This statistic is a normalised measure in the interval [0,1]. It is equal to zero if $L(\beta^{ML})$ is equal to $L(0)$, i.e. the model has no explanatory capability; it is equal to one if the model gives a probability equal to one of observing the choices actually made by each user in the sample, i.e. the model has perfect capability to reproduce observed choices.

3 SP and RP data

Model estimation (specification-calibration-validation cycle) can be performed starting from information on the behaviour of each user of a sample. This approach is called disaggregate estimation of models. The surveys used to gather elementary information might belong to two different classes: surveys relative to the actual behaviour in a real context (*Revealed Preferences* or *RP* surveys) or surveys relative to the hypothetical behaviour in fictitious scenarios (*Stated Preference* or *SP* surveys). These surveys provide information on residents' choice relevant for the model to be calibrated. Stated preferences (SP) surveys differ in that they are conceptually equivalent to a laboratory experiment designed with a larger number of "degrees of freedom" very similar to the market surveys.

In applications, a wide range of possible surveys to collect data may be considered. Examples are home interviews, postal or telephone interviews, road side surveys, number-plate surveys, on board interviews or procedures combining several of these techniques.

It is, in general, necessary to survey many types of individuals in order to obtain representative results. If we sample randomly, we may require large samples in order to achieve an adequate number of observations about minority choices.

During the interview, the decision-maker is usually presented with different scenarios or choice contexts. The decision-maker can be asked about different types of preference:

choice, i.e. an indication of which option he/she would choose in that context;

ranking, i.e. a ranking of the available options according to his/her preferences;

rating, i.e. the assignment of a vote of preference on a predefined scale for each alternative option.

Sample design aims at ensuring that the data to be examined provide the greatest amount of useful information about the population of interest at the lowest possible cost; the problem remains of how to use the data (i.e. expand the values in the sample) in order to make correct inferences. Thus two difficulties exist:

- how to ensure a representative sample;
- how to extract valid conclusions from a sample satisfying the above condition.

Unfortunately, there are no straightforward and objective answers to the calculation of sample size in every situation. This happens because many of their inputs are relatively subjective and uncertain; therefore they must be produced by the analyst after careful consideration of the problem.

SP surveys have several advantages over RP surveys, which can be summarised as follows:

- they allow the introduction of choice alternatives not available at the time of the survey ;
- they can control the variation of relevant attributes outside the present range to obtain better estimates of the relative coefficients;
- they can introduce new attributes not present in the real choice context;
- they can collect more information, i.e. larger samples, per unit cost since each interviewee is usually asked about several choice contexts.

These advantages are obtained at a price of introducing some

distortion in the results and in the models calibrated. Assuming the SP survey well designed and employed, distortions stem from the possible differences between stated and real choice behaviour; if the user experienced a real situation, his/her behaviour might be different from that stated during the SP survey. Differences in behaviour may be due to several factors. For example, the context suggested might be or appear to be unrealistic, some attributes of the suggested alternative relevant for the decision maker might be missing, there may be fatigue and justification bias effects. However, it should be noted that some of these problems are structural, or ingrained in SP surveys technique, while others can be solved by careful design and execution of the surveys bringing them as close as possible to real choice contexts.

On the other hand, RP data give an indication of the actual choices made by households and this data can be used to estimate the parameters of any households' location choice model. They have a high degree of reliability in that they represent actual behaviour. In RP studies choice-base sampling could provide a very cost-efficient method; this is made more interesting by the fact that a simple correction can usually be made in model estimation to avoid statistical bias. However RP data describe the compromises households make, not their true preferences. The disequilibrium and habit that affects real world residential location behaviour cause households to not necessarily realise their preferences, but rather to stay put or accept what market has to offer. A related problem is the existence of correlation among the attributes in real world data. For example, a positive correlation is to be expected between house size and travel time to work in North America cities because larger houses tend to be located towards the edges of built up areas. Such correlations make it difficult to separate the influences of different factors using statistical analyses of revealed preference data. In addition, collecting real world data is usually very expensive and time consuming.

From the above, it is clear that SP surveys, in spite of their considerable application potential, should be seen as complementary, rather than alternative, to RP techniques. The advantages and disadvantages of the two techniques compensate each other and the techniques can be used jointly to build demand models.

Experimental evidence indicates that the combined use of RP and SP data for estimating the parameters usually results in an improvement in statistical precision and in more reasonable parameter values.

4 The case of Calgary

The research described here was done to investigate the potential impacts of light rail transit (LRT) development in Calgary (Canada) on land values in general and residential land values in particular (Hunt et. al., 1993). Descriptions of the hypothetical alternatives considered in the stated preference experiments performed for this research were developed by selecting one out of a set of possible values for each attribute influencing housing preferences and combining these selected values into a bundle representing a complete alternative. With 3 possible values for one attribute and 2 for each of the other four, the total number of distinct hypothetical alternatives was 48. Neither fractional design nor dominance analysis was taken into account to efficiently reduce the number of alternatives.

In 1992 over 390 choice experiments were conducted with individuals selected randomly at various shopping centres in Calgary. Each experiment was a voluntary interview where the respondent was approached and asked to rank four hypothetical housing alternatives in order of preference from best to worst, taking into account the needs and wants of the respondent's present household. In each case these four alternatives were selected randomly from the full set of 48 alternatives in the deck of cards. Each respondent was also asked a variety of questions regarding socio-economic status and household characteristics, including the following:

- home location
- workplace location, if working
- number of people in household
- combined of licensed drivers in household
- number of cars available for use by people in household.

The result was a data set with 377 disaggregate stated preference observations. This data set was used to estimate the coefficients of a Logit model reported in the following.

It is not unreasonable to expect that those people living within walking distance of LRT and those people not living within walking distance of LRT will differ in terms of their perceptions of the benefits of proximity to LRT. This is because there will be some self-selection in that households most concerned about being close to LRT will be more inclined to move to locations close to LRT. As time progresses this will lead to a relatively larger proportion of LRT-proximity-sensitive households in areas close to the LRT. Those households who actually live close to LRT will have had relatively more opportunity to use LRT to its full advantage and may thereby develop a more accurate appreciation of the actual (either increased or decreased) value of being within walking distance of the service. The results for a utility function that distinguishes between the perceptions of those who do and those who do not live within walking distance of LRT are reported in table 1. The factors found to have influence in residential location choice behaviour in Calgary are in the following reported.

$COST_i$ = money cost per month for alternative i;
INC = the total annual income for the respondent's household, in Canadian dollars per year;
$BEDS^{2-}_i$ = number of bedrooms for alternative i when number of persons in household is 2 or less and 0 when number of persons in household is more than 2;
$BEDS^{3+}_i$ = number of bedrooms for alternative i when number of persons in household is 3 or more and 0 when number of persons in household is less than 3;

$WORK_i$ = in-vehicle travel time for trip from alternative *i* to workplace, in minutes;

$SHOP_i$ = in-vehicle travel time for trip from alternative *i* to shopping centre, in minutes;

$LRTP^C_i$ = 1 when an LRT station is within walking distance of alternative *i* and the respondent's actual home location is within 400 metres walking distance of an LRT station (designated "C" for close) and 0 otherwise;

$LRTP^F_i$ = 1 when an LRT station is within walking distance of alternative *i* and the respondent's actual home location is not within 400 metres walking distance of an LRT station (designated "F" for far); and 0 otherwise.

ρ^2=0.236	$COST_i$ / INC	$BEDS^{2-}_i$	$BEDS^{3+}_i$	$WORK_i$	$SHOP_i$	$LRTP^C_i$	$LRTP^F_i$
β	-125.2	0.1279	0.8658	-0.05571	-0.02981	1.369	0.5952
t-ratio	-12.7	2.1	12.0	-9.6	-3.5	3.9	6.8

Table 1: calibration results

All the coefficients estimates are statistically significant and have signs consistent with what would be expected. The value ρ^2 for is higher for any other function, indicating that this utility function provides the best model fit out of those considered. The t-statistic for the difference between the coefficient estimates for $LRTP^C_i$ and $LRTP^F_i$ is 2.13, which means that these two variables should kept separate. The ratio between these two estimates indicates that being within walking distance of LRT is 2.30 times as important to households located within walking distance of LRT in reality. It should be noted that only 10% of those interviewed were from households located within walking distance of LRT. This will have increased the sampling error for the information concerning these households' evaluations of proximity to LRT in particular. The amount of confidence placed in the coefficient estimate for $LRTP^C_i$ must be reduced accordingly.

Nevertheless, the results indicate that there is a statistically significant difference in these two groups' perceptions of the importance of being within walking distance of LRT. Several studies suggest that there tends to be at least two groups of households: one group that tends to use public transport and for whom public transport service is an important factor influencing the quality of residential locations; and another group that tends not to use public transport and for whom public transport service is almost irrelevant to the quality of housing locations. One or the other of these groups can dominate a given sample of housing preference observations, and when this happens the overall indications from the sample are swayed accordingly. Moreover, the distance from workplace zone weights twice the distance from shop centres. This confirms the wish of people to choose a location as close as possible to workplace.

5 The case of Naples

The problem of simulating the clearing of an urban residential market can be described as that of allocating *w* types of residents in *i* sub-areas or zones of the city from *j* places of employment. In this case w refers to the income level; two groups are identified: a high/medium and a low income groups (Cascetta et al., 2000).

The inputs to the calibration of the model include direct observations data. These inputs are of two different types and they are Population Census 1991 (ISTAT, 1991) and a mobility survey employed by the company ITER (1998). The information got from the latter are mainly related to transport rather than to land-use. The study area is the urban area of the city of Naples.

The utility function parameters have been estimated with the Alogit software package for the residence choice location model. The calibration results are reported in tables 2 and 3 and are significant with a sample of 485 interviewees.

The attributes of the systematic utility are in the following reported:

lnSTOCK (i) = ln of the housing stock in zone *i*;

PRICE (i) = price per square meter of the houses in zone *i*;

ACA_SER^c(i) = active accessibility to services in zone *i* for residents of type *c*;

$Y^c_{work}(i,j)$ = logsum of the mode choice *m* between *i* and j for work purpose and type *c* of residents;

PREST (i) = 1 if zone *i* is prestigious; and 0 otherwise;

CH (i) = council houses ratio in zone *i*.

ρ^2=0.718	lnSTOCK (i)	PRICE (i)	ACA_SER^c(i)	$Y^c_{work}(i,j)$	PREST(i)	CH(i)
β	0.2106	-0.049	0.945	2.578	0.3894	-1.365
t-ratio	4.1	-1.3	1.6	11.1	2.5	-3.1

Table 2: calibration results for the high/medium income group

ρ^2=0.1102	lnSTOCK (i)	PRICE (i)	$Y^c_{work}(i,j)$	PREST(i)
β	0.2806	-0.385	3.698	-2.2487
t-ratio	2.0	-2.0	5.5	-0.4

Table 3: calibration results for the low income group

The β parameters are all of the expected sign. Particularly significant are the transportation attributes, which show the relevance of the transportation supply and the level-of-service among the various zones. The housing stock seems to weight similarly for the two income classes, the other variables are highly different. As a matter of fact, high/medium income residents are willing to pay more for a more accessible and prestigious zone. Such considerations are confirmed by the positive and high values of the logsum and prestige parameters and by the negativity of the council house ratio parameter. On the other hand, low income people do not perceive the same value, but give a great relevance to the transport level-of-service, giving them the possibility to easily leave their zone for work purpose. Thus, prestige zone parameter has a very negative value for the low income

group. The PRICE is not significant for high/medium income group, while it is for the low one; on the other hand the STOCK attribute has almost the same weight for both groups as it was already above stressed.
The low significance of location variables (CENTRE, FIRST RING and SECOND RING), which led to their exclusion from the utility function, can be due to the accessibility, transport and prestige variables, which seem to be related to the location of the zone. The low ρ^2 value for the low income group is mainly due to the very small sample.

6 Comparison and conclusions

A comparison between the two cases should be done. The importance of the transport component is stressed in both studies, especially in the case of Naples with the introduction of the logsum of the mode choice. The classification of residents into income groups has turned out to be a good method compared to that used in the case of Calgary where respondents were classified according to the distance to an LRT station. Furthermore, it should be stressed that even if the RP survey was employed to analyse the mobility system in the urban area of the city of Naples, it has given again good results also for the investigation of residential location choice behaviour.
Various housing attributes and household characteristics have been shown to have a statistically significant influence on housing preferences. This includes several transportation-related attributes, thus indicating that the transport system has an effect on the attractiveness and hence on the value of residential location both in Calgary and in Naples. The stated preference techniques that were used, in the case of Calgary, were found to be very successful. It has provided necessary tools for planning analysis and has contributed to the further understanding of the behaviour of the urban system in Calgary. On the other hand, the revealed preference techniques has given again good results for the investigation of residential location choice behaviour. Even if the sample was very small for the low income group, the calibration results were significant and all of the expected sign. Further work should consider for the area of Naples a survey aimed at analysing residential location choice. In this case, direct questions to the Neapolitan residents on their housing preferences should be asked.

References

CASCETTA, E. (2001) *Transportation systems engineering: theory and methods*. Kluwer.

CASCETTA, E, BIGGIERO L AND PAGLIARA F (2000) An integrated land-use/transport model calibrated on the urban are of the city of Naples, submitted to the *8th Euro Working Group meeting*, Rome, Italy.

HUNT, J. D., ABRAHAM, J. E. AND MCMILLAN J. D. P. (1993) Factors influencing residential location choice behaviour in Calgary Canada submitted for presentation as a full paper and for publication in the proceeding of the 1993 Annual Conference of the Institute of Transportation Engineers held in Edmonton, Canada 6-9 June.

ISTAT 1991 Census data.

ITER 1998 RP survey.

WEGNER, M. AND FURST, F. (1999) Land-use and transport interaction: State of the art, final draft of the TRANSLAND project funded by the European Commission within the 4th FP, Institute of Spatial University of Dortmund, Germany.

4 Concluding question

Finally the question arises: have the manufacturers taken into consideration the consequences of the stoppage in fuel supply from one wing tank for stability and control of their aircrafts? Is the magnitude of the ΔI of the possible change of *LILDW* analyzed and based on that analysis is it ensured that the landing limits in the balance graphs are laid down with an adequate safety reserve, i.e. is it ensured that the intersection of *LDW* and the actual (changed) *LILDW* in the most inauspicious case will not be positioned out of the actually safe area?

General references

[1] Duncan, W.J., *Control and Stability of Aircraft*, University Press, Cambridge, 1972.

[2] Mhitaryan, A.M., *Flight Dynamics*, Mashinostroenie, Moscow, 1978. (in Russian)

[3] Richter, W. (ed.), *Luftfahrzeuge*, in: Fachwissen des Ingenieurs, Bd. 8,VEB Fachbuchverlag, Leipzig, 1986.

[4] Airport Handling Manual, IATA, 1996.