Strategies for specification search as a cause of bias and inaccuracy of parameter estimates

An important problem in model development, illustrated by MonteCarlo simulation and Bootstrap

Dr Karin Brundell-Freij
Department of Technology and Society, Lund University, Sweden

1. Introduction

As we all know, model predictions and estimation results are erroneous. What we do not know is the size of those errors (had we done so, we would have compensated for them). Never the less, professionalism requires that we couple the model results we use, with some kind of measure of their estimated quality. Accuracy thus is a key issue of modelling – not only to *obtain*, but to properly *describe*.

The typical tool we would use for description of model accuracy is standard errors of obtained parameter estimates. However, standard errors are only designed to illustrate a smaller part of those model errors that may arise from the complex process of developing a transport model. This paper discusses and investigates some such limitations in relation to the errors that may arise from specification search[1]. Initial analyses, based on a combination of a real data set and simulation tools, will show that there may be considerable inaccuracy and bias caused by systematic factors that are outside the scope of standard error.

Despite the fact that the concrete example, and implicitly much of the discussion, relates to models for discrete choice applied to transport demand, the general conclusions would apply also to a vast range of other types of modelling.

2. Model development as a process

Classic statistical theory draws a sharp line between the specification of a model - which basically is supposed to be done *ex ante*, based on correct assumptions about the basic processes in the population, on one hand – and the estimation of its parameters, which is supposed to be done *ex post*, based on observed data. See figure 1.

The image of figure 1 does however fit very badly with the process in which real transport models (in fact, most real models) are developed. The final result here may rather be seen as a *combination* of

- a set of a priori assumptions about behavioural principles, important variables and desired model properties, and

[1] See Brundell-Freij (2000) for a somewhat more elaborated discussion on different error types.

- a collected data set (normally cross-sectional) coupling transport demand with observations on potential input variables

which communicate with each other on a more equal basis.

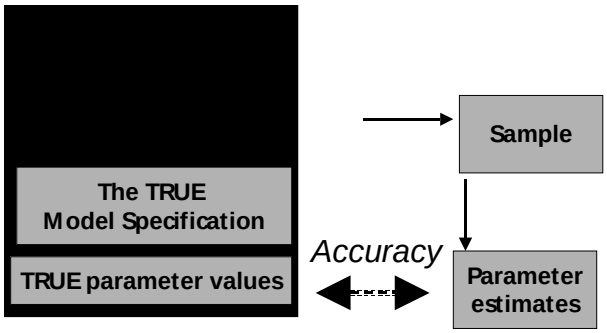

Figure 1 Model development in statistical theory

In this, ”real world”, modelling, a set of model properties that were regarded as reasonable *ex ante* (e.g. intended segmentation or the inclusion of specific important policy variables) may be traded-off against each other and against model fit, based on feed-back from numerous test and re-test estimations of a large set of potential model specifications. This process is tedious, but certainly necessary to obtain a useful final result. An attempt to visualize the process is presented in figure 2.

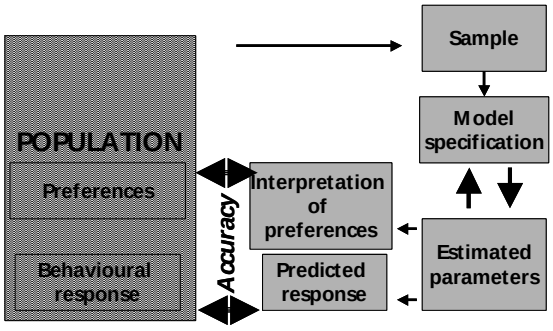

Figure 2 Model development in reality

The model development process described above (“numerous test and re-test…”) is indicated by the feed-back loop between the specification, and parameter boxes in figure 2.

Another important difference between figure 1 and 2, is that in ‘reality’, the notion of a ‘true’ model is (most probably) irrelevant. From this follows, that whatever model we define, not even in theory are there any ‘true targets’ for the parameter estimates to be evaluated against.

But how, then, may model quality be defined? A reasonable approach is that the merit of a model is given by the differences between some interesting properties of the real population, on one hand, and what we think we learn about those properties, based on our interpretation of the estimated model, on the other. This approach is illustrated by the indicators “interpretation of preferences” and “predicted response”, in figure 2.

But let us go back to the feed-back between specification and estimation stages. Much of this process is un-standardised, and based on the experience and 'tacit knowledge' of the model developer. But also standardised tools for choosing proper model specifications, based on estimation results, has been presented in literature. For discrete choice models, Ben-Akiva and Lerman (1985) mention e.g. quasi-t-tests (both for deciding about the inclusion of a potentially non-contributing variable, and for judging about generic vs. segment-specific parameters), Likelihood-ratio tests (for choosing between restricted and less restricted model formulations) and the rho-2 measure for indicating model fit. These tests are formed to choose between different, discrete, model specifications that are pre-defined by the model developer.

One step further is taken by Bhat (1996) who suggests a method for "automatic" definition of a suitable segmentation of the sample with respect to the assessment of certain variables. Sørensen (2001) suggests another standardised, but non-parametric, approach to the same segmentation problem, while Linveld (2001) suggests an automatic method for selecting optimal cut-points for a piece-wise linear utility function. In those latter examples, the potential number of competing model specifications is (almost) infinite. Anyway, it is clear that in many cases, the model specification finally selected will relate closely to initial estimation results based on the same data.

When model development is regarded as a process in communication with the data in this way, the classical – frequentist - perspective of hypotheses testing etc becomes less applicable. More fruitful alternatives are then offered by Baysian approaches (as discussed in Geisser (1993)). Also within those, however, estimates of *accuracy* (now regarded as measures of certainty) are key indicators.

3 Standard errors of parameters as a quality indicator

Two well-known standard tools for estimating quality of estimated parameters are offered traditionally – bias and standard error.

Note that neither estimates of bias, nor standard error estimates, directly refer to the quality of *the specific* estimated model. In single cases, also estimates with large standard error may be very close to the corresponding true value. Rather, standard error and bias may be regarded as estimates of the quality of the *method* that was used to produce the estimates, which in turn (in a Baysian interpretation) forms the basis for our trust in the estimates.

From the way standard errors are defined, however, it follows that they only measure the quality of certain aspects of the method. As standard errors of estimated regression parameters (including those that are ML estimated, e.g. logit parameters) are defined, they indicate the variation that you would expect

- if the model specification was correct (‘*under the model’*)
- if the attributes were kept constant (‘*conditioned on attribute values’*)
- and you repeatedly resampled the random components of the model
- each time estimating the same parameters with the specific method used

This, hypothetical, model is illustrated in figure 3.

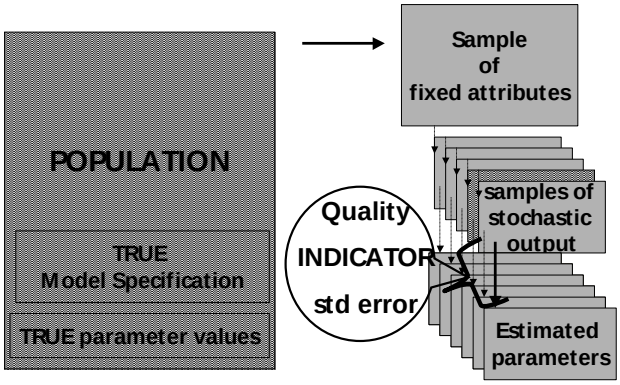

Figure 3 Standard error as a result of hypothetical resampling.

Now, taking a broader view on model/method quality, it becomes clear that there are other sources of uncertainty affecting estimates, on top of that our sample may not correctly represent the overall distribution of the random components.

As was stated above, standard errors of direct parameter estimates are the most often presented indicator of model quality. It has, however, been argued for good reasons (e.g. by Munizaga et al (2000)) that quality of such parameter estimates are far from ideal indicators of model quality, since these estimates as such are not ‘final’ output. The final output is argued to rather be the predictions (reactions to policy measures) that the model will produce. In this view, we may well live with inaccurate parameters, as long as they do not lead to inaccurate model predictions. This corresponds with the modified criteria, based on “predicted response”, which was introduced in figure 2.

In this paper, we take an alternative approach in using the quality of *indirect* parameter estimates (such as values-of time) as key indicators of model quality. This is for two reasons: Firstly, such estimates are often used directly for behavioural interpretation and in assessment. Thus, they represent “final” output in another of the senses introduced in figure 2: they form our understanding of basic preferences in the population. Secondly, accuracy of model predictions will (naturally) be very sensitive to the policies imposed. By evaluating values-of-time rather than predictions, we avoid a crucial, but somewhat arbitrary, selection of policies for which to evaluate model performance.

4. Model specification as a source of inaccuracy

It has long been observed and discussed in literature that model formulation influences model results. Both the effect of varying overall model assumptions (e.g. Ortuzar and Garrido (1998)) and more detailed model specification (e.g. Gaudry et. al (1989)) on model output, has been investigated. Many more authors (e.g. Gärling (1994) and Brög and Erl (1983)) have raised doubts to the applicability of estimated models, since the model assumptions behind them have been assumed to be erroneous, which has been assumed to lead to incorrect parameter estimates. In opposition to this, however, e.g. Munizaga et al (2000) and Brundell-Freij (1996) has found that detailed specification may not in all cases be of major importance for model outcome. The overall conclusion, however, from many of those observations seem to have been that it is very important to choose the 'correct' model, because otherwise the estimation results may not be trusted.

But, referring again to figure 2 above, it seems more reasonable to accept that *all* model specifications are necessarily erroneous. In this light, the variation of model estimates (direct or indirect) with varying model definition may be regarded as a parallel to the variation with different samples of random components. From such variation, we would not typically draw the conclusion that "not all estimates can be correct, so we have to find out which is the correct one". Rather, we would understand the variation as contributing to our overall uncertainty about parameter estimates.

In the same way, if estimates are similar, irrespective of model specification, then we may regard them as certain. If, on the other hand, they vary much depending on specification, then we should feel less certain about them.

Much of the final model specification may be regarded as an outcome of model estimations (through a more or less specified search process, see section 2 above). In this meaning, the part of variability of estimates that is attributable to varying model specification is linked to sampling probabilities, in the same way as ordinary standard error is. (Since the relation is complex, it is however most unlikely that it would be possible to compute a closed form estimate for the distribution.)

For this distribution of estimates after model selection (as opposed to 'for a specific model'), we may raise the usual questions about variability and potential bias. These problems have been given some interest within the field of statistics (see e.g. Hjorth(1992)), but is generally very little discussed. One may suspect that estimates after model selection have a different distribution than they have for a fixed model. In some simple, but realistic, cases it can also easily be shown and understood that parameter estimates after model selection generally have a distorted distribution[2]. In real cases, with complex model search mechanisms and

[2] A simple example: A commonly applied strategy is to reestimate regression models, excluding variables the parameters of which in an initial, 'complete' estimation has turned out non-significant (based on e.g. a quasi-T-test). This strategy will, after model selection, make parameter estimates infinitely close to zero very improbable,

covariance matrices of data, it is harder to predetermine the effect on parameter accuracy from model search.

5 The test model

5.1 Base data

The analyses in the following are based on a real data set, collected within a commuter survey covering the corridor between the cities of Lund and Malmoe, 20 kilometres apart in the south-west of Sweden.

Four possible modes were defined (car, car pooling, train and bus). 845 observations were collected through choice based sampling, and trip characteristics for hypothetical trip alternatives were assigned objectively for each available mode.

Although the following analyses is based on hypothetical techniques (Simulation and Bootstrap), it has been ensured throughout that the multidimensional distribution of attribute values was kept as it was observed in the real study. This is essential, since much of the limitations on model quality is set by the attribute covariance matrix in the population, which thus has to be represented realistically. At the same time, however, this means that whatever specific conclusions that are drawn from the analyses, they are primarily applicable to the population and modelling task represented in this study.

5.2 Specification search algorithm

One of the main aims of the analyses is to investigate the inaccuracy induced by the specification search. To do this in a standardised way, we have to formulate an explicit, realistic algorithm for such search. This strategy has to be standardised, despite the fact that in real modelling specification search is typically based on a combination of experience, 'gut feeling' and evaluation of quality indicators. In this case the strategy that was formulated was based on introspection (What would I normally do?), while also considering experiences from professional discussions with other model developers (What have they said that they think you should do?).

Thus, the aim in general was to – to the extent possible - directly formalise a realistic process of specification search. The formalised process as it was finally defined was however somewhat a compromise between realism and practicality of the programming task. The following basic limitations were set (probably rather restrictive in comparison to what is performed in a real case):

---

because such values are seldom significant. It will be much more probable that our final model states that the variable has no influence (i.e. that the final 'parameter' is exactly zero), since this will be the end result if the first estimation does not pass the quasi-T-test. Thus, the distribution of final estimates will be discontinous.

- a Multi Nominal Logit structure was regarded as given
- a set of variables (table 1) to be included in all tested models was defined and their assigned values were regarded as given
- socio-economic segmentation was only tested for gender, and tested only in three distinct forms:
  - for each alternative-specific constant
  - for each time component
  - or cost

Table 1 Variables in the different model specifications

	optionally specified to	optionally segmented for
In vehicle-time	car / regional p.t / local p.t	gender
Walk/bicycle time		
Adaptation time	waiting / transfer	
Cost		gender
Alternative-specific constants		gender

The defined set of variables and the restrictions formulated above, pointed out 32 (2^5) model specifications to estimate. The choice of a final model between them was formalised as being based on the following formalised priorities:

- all parameters for time(s) and cost had to have intuitively correct signs for a model to be accepted
- if at least one model could be estimated that had all t-values for time and cost components >1.5), other models would not be considered
- increased segmentation and specificity with respect to preferences was generally regarded as an improvement, but had never the less to be justified by significant (quasi-t-test) difference between parameter estimates applying to different segments/modes/time types
- in a choice between gender segmentation and mode-specific values of in-vehicle time respectively, the former was preferred.

In cases when these rules did not point out one specific estimated model as being the best, the final choice was based on adjusted rho-2 values according to Ben-Akiva and Lerman (1985).

The formalised model selection procedure above was implemented in computerised form in a MATLAB environment. The developed tool combined random regeneration of data (see section 5.4), external ALOGIT estimation of the 32 optional models and automatic evaluation of estimation output according to the defined strategy for specification search.

Obviously the model selection procedure implemented in this work should not be regarded an 'optimal' one. Hopefully, however, it is realistic enough for the results to obtain a certain level of generality.

5.3 Base model

Applying the model development strategy formulated in section 5.2 on the observed base data, we arrived at the restricted model presented in table 2.

Table 2 also introduces the indirect parameters (value-of-time estimates) that will be used in the following to illustrate model accuracy. These were the same for all models, and refer respectively to

- male in-vehicle time for travel with regional public transport
- male waiting time [3]

Table 2 Base model, resulting from specification search, estimated on base data

	specified to	**estimated parameter (estimated s.e.)**	
In vehicle-time (min)		-0.127	(0.014)
Walk/bicycle time (min)		-0.041	(0.017)
Adaptation time (min)	waiting	-0.083	(0.022)
	transfer	-0.048	(0.023)
Cost (SEK)		-0.096	(0.021)
Alternative-specific constants			
car pool		-0.664	
train		-0.128	
regional bus		0.65	
Indirect parameter estimates			
Male VOT reg p.t. (SEK/h)		78.70	(13.70)
Male VOT waiting (SEK/h)		51.70	(5.10)

[3] For this specific model formulation, we see that there is no differentiation between neither gender, nor modes when it comes to the assessment of these time components. For comparable measures to be computable also for model specifications where such differentiation is made, however, the specification to mode and gender is necessary

5.4 Monte Carlo simulations and Bootstrap

Modern development within the field of statistics has been closely related to the development of computer technology. Many of the problems that statisticians have been struggling with for a long time can now easily be approached with the use of computer based methods, see Hjorth (1992) for a general discussion. Despite the general usefulness of such methods, the results may be questioned, since they rely heavily on the input assumptions, see eg (Daly et al (2002), Tuleda (2000)). Thus, the analyst basically control the results by defining those assumptions, which may often be chosen rather freely. Keeping those limitations in mind, two different simulation based methods are used to serve different purposes in the analyses in this paper.

Monte Carlo simulation is a straight-forward method for investigating the distribution of a function (in our case an estimator) of one or more random variables following a known distribution. It has been widely used for the analyses of the behaviour of travel demand model. Both Tudela (2000) and Sola Conde and Daly (2002) does however point to the fact that the method is very sensitive to input assumptions, and therefore potentially less useful for drawing general conclusions, than it first may seem.

In this paper, Monte Carlo simulations were used for analysing the effect of resampling of random model components under a specified data generation process (consistent with the MNL base model estimated in table 2), to investigate the influence of such variation on estimation output and specification selection.

Attributes for each observation were kept constant throughout simulations, and only choices were regenerated, based on stochastic utility. The results thus illustrate the variability of the estimates *conditioned on attribute values, under the base model.* This approach relates closely to the hypothetical model in figure 3, which underpins the concept of standard error. Two separate analyses were made, one in which the model specification was kept constant, and another where the distribution of estimates *after model selection,* was studied.

Bootstrap. While Monte Carlo methods require assumptions or knowledge about the distribution of random components, Bootstrap (in its original form) is free from such assumptions. The Bootstrap method[4] approximates the distribution of a given estimator θ over different samples from the population, F, by resampling observations from the *sample* distribution, F_n instead.

In this work, Bootstrap was used to resample observations (attributes and choices) to illustrate the variability that arises from the whole chain of model imperfections, sampling and estimation. Two separate analyses were conducted: One in which the model specification was

[4] see Efron(1993) for a more formal and thorough discussion

kept constant, and another in which the variability after specification search was studied. Generally, the sample size was 845 in all analyses (as they were in the original sample).

The separate analyses performed may thus be summarised as in table 3.

Table 3 The different analyses performed, and their notation in the following

	Variability for base model	**Variability *after model selection***
Resampling of random components, data generating process = base model	***MCbase*** Monte Carlo N=845	***MCsel*** Monte Carlo N=845
Resampling of whole observations from F_n	***Bootbase*** Bootstrap N=845	***Bootsel*** Bootstrap N=845

To make it possible to analyse estimate distribution, each separate analysis comprised 100 full sets of

- regeneration of data
- estimation of model(s)
- model selection (if applicable).

5.5 Analysis

Table 4 summarises the average estimates of the indicators for the different analyses.

Table 4 Average indicator estimates for different analyses. * indicates significant difference (t-test) attributable to specification search.

	MCbase	MCsel (all models)	Bootbase	Bootsel (all models)	Base model (reference)
Male VOT reg pt	79.86	123.17*a)	80.62	86.82*	78.69
(s.e.)	(1.98)	(23.06)	(2.32)	(2.79)	(13.69)
Male VOT waiting	40.41	40.42	25.69	44.63*	51.69
(s.e)	(1.03)	(1.80)	(1.39)	(1.63)	(5.10)

a) strongly influenced by single outlier. Average without outlier: 106.36. s.e.= 15.95

The results in table 4 strongly supports the initial assumption that the process of specification search (based on estimates) may introduce bias-by-selection in obtained model parameters.

Our imposed strategies for specifiaction clearly seems to favour models with large estimates of value-of- time in this case[5].

Further, the comparison between Monte Carlo simulations and Bootstrap results, indicate that there are 'imperfections' (as regarded by the model) in real behaviour that actually helps stabilising the model selection procedure in comparison to what would be expected under the model.

At least for the Monte Carlo estimations, it is clear that the base model parameters for the initial sample (rightmost values in table 4) are target value for the estimations. These parameter values have been the basis for the data generating process. Never the less, one of the estimated values-of-time are, on average after model selection, more than 25% higher than that target value. This is partly because value-of-time estimates (being ratio estimates) are biased[6], but also a result of the specification search process (cf. the average VOT under the model, "MCbase" with the corresponding value after model selection "MCsel").

The most important conclusion however, is probably that model selection adds variability. The standard error of the base model parameters should, theoretically, represent the standard *deviation* among repeated estimations. That would correspond to the standard *errors* for those averages (as presented in table 4), multiplied by a factor 10. Both the Bootstrap, and certainly the Monte Carlo illustrations indicate that the variability after model selection is far larger than what would be assumed from the base model estimate.

Often, the specification search (naturally) ends with the same model specification that was selected for the base data. For the Bootstrap simulations that was the case for 62 of the 100 runs, for the Monte Carlo 42 of the 100 model selection procedures stopped at the base specification. With this in mind, the differences between estimates for a constant specified model, and the corresponding estimates after model selection, must be regarded as surprisingly large. The reasons behind those differences are however not easily disclosed, since the process is complex. Some indications may however be obtained by a closer look at the distribution of parameter estimates in different cases.

One way of doing this, is to compare what was found to be the *overall* distribution of parameter estimates (Bootstrap) for the 100 versions of the base model, with the parameter estimates that result from the 100 model specifications resulting by the selection procedure. This comparison is made in figure 4 below.

[5] For Monte Carlo ans well as Bootstrap analyses, and for both in-vehicle time and waiting time, average VOT after model selection is higher than the corresponding value for the base model specification. For three of the four comparisons, the difference is statistically significant.

[6] One of the reasons why Jack-knife methods have been much applied for value-of-time estimation is that they may reduce such bias..

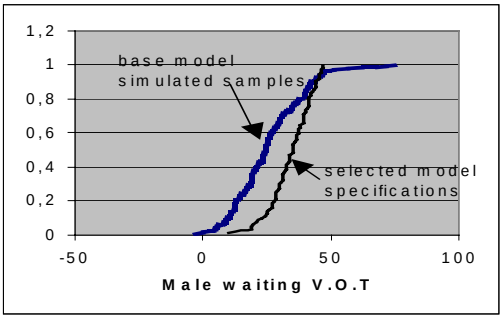

Figure 4. Distribution of estimates (VOT male waiting time) obtained by the base model in all bootstrap replicates, and for the selected model specifications, respectively

From figure 4 we understand that whenever estimates of the base model is in the lower end of the distribution, other model specifications are selected. (Remember that in 62 of the 100 cases, the selected specification is the base model.)

This bias-by-selection obviously is a very strong effect, despite the fact that none of the selection mechanisms introduced does explicitely favour large estimates of value-of-time. The result in figure 4 thus has to be an effect of more subtle processes, arising, for example, from the correlation between different parameter estimates.

6 Conclusions

Model selection procedures are applied for good reasons, to avoid mis-specifications that would increase the risk of inefficient policy making. However, such procedures do increase the risk of biasing the outcome of the estimation procedure through bias-by-selection. The initial analyses of this issue that were conducted in this paper showed that substantial bias, and increased variability, may be introduced already by quite restricted specification search.

But it was also shown that by the use of modern statistical techniques, it is possible to improve the description of model quality, and incorporate sources of error that has not been captured according to current standards.

ACKNOWLEDGEMENTS

This is a modified version of a paper that was initially presented at European Transport Conference in Cambridge in 2000, Brundell-Freij (2000). The introductory discussion in that paper covers some additional aspects on sampling and estimation, excluded here. The reason for this modification was to leave room for a (hopefully) improved didactic approach in this later version.

The work was initialised under a research grant by the Swedish National Transportation Research Board, and was inspired by previous work by professor Urban Hjorth, Gothenburg. The support from both is gratefully acknowledged.

BIBLIOGRAPHY

ALOGIT software. Hague Consulting group, The Netherland

Ben-Akiva, M and Lerman, S (1985) *Discrete Choice Analysis*. MIT Press, Cambridge, Massachusetts

Bhat, C (1996) *An Endogenous Segmentation Mode Choice Model with Application to Intercity Travel*. Transportation Research Board 75th Annual Meeting, Washington

Brundell-Freij, K. (1996) *How Good is an Estimated Logit Model?* Proceedings from PTRC Annual Conference, PTRC, London.

Brundell-Freij , K. (2000) *Sampling, specification and estimation as sources of inaccuracy in complex transport models*. Proceedings from European Transport Conference 2001, Cambridge. PTRC, London.

Brög, W. , Erl, W. (1983) *Application of a model for individual behaviour to explain houshold activity patterns in urban area and to forecast behavioural changes.* In: recent advances in Travel Demand Analysis, Eds: Susan Carpenter, Peter Jones. Gower Publishing.

Efron, B., Tibshirani, R. (1993) *An introduction to the Bootstrap.* Chapman & Hall, New York.

Gaudry, M.J.I, Jara-Diaz, S.R., Ortuzar J.d.D (1989) *Value-of-time sensitivity to model specification*. Transportation research 23B, No 2. Pergamon

Geisser, S (1993) *Predictive Inference: An Introduction*. Chapman & Hall, New York

Gärling, T.(1994) *Behavioral Assumptions Overlooked in Travel-Choice Modelling.* Paper presented at the 7th international conference on travel behaviour. Valle, Chile.

Hjorth, U. (1992) *Computer Intensive Statistical Methods*. Chapman and Hall, London.

Linveld, C.(2001) *Non-linear utility functions in MNL discrete-choice models*. Proceedings from European Transport Conference 2001, Cambridge. PTRC, London

MATLAB software (1994). The Math Works Inc.

Munizaga, M., Heydecker, B.G., Ortuzar, J. d. D. (2000) *Representation of heteroskedasticity in discrete choice models*. Transportation Reasearch Volume 34 B, No 3. Pergamon.

Ortuzar, J.d.D, Garrido R.A. (1998). *Methodological Developments*. Workshop report. Eight International conference on Travel Behaviour. Austin, Texas.

Sola Conde, P., Daly, A. (2002) *Person specific models in SP analyses.* Proceedings from European Transport Conference 2002, Cambridge. PTRC, London.

Sørensen, M.V. (2001) *An alternative data segmentation method*. Proceedings from European Transport Conference 2001, Cambridge. PTRC, London.

Tudela, A.M. (2000) *Simulation as a necessary step in the design of Stated Preference experiments*. Proceedings from European Transport Conference 2000, Cambridge. PTRC, London.