Monograph:
Ortuzar, J. de D. (1980) Multimodal Choice Modelling – Some Relevant Issues. Working Paper. Institute of Transport Studies, University of Leeds , Leeds, UK.

Working Paper 138

ABSTRACT

ORTUZAR, J. de D. (1980) Multimodal choice modelling - some relevant issues. Leeds: University of Leeds, Inst. Transp. Stud., WP 138. (unpublished)

This paper gives an overview of the most relevant issues relating to the application of multimodal choice models ranging from data considerations, such as alternative sampling strategies and measurement techniques, to the hotly debated aggregation issue. Particular emphasis is placed on the specification and estimation problems of disaggregate choice models.

Dr. Ortuzar's address is: Departamento de Ingenieria de Transporte
Universidad Catolica de Chile
Casilla 114-D
Santiago - Chile.

CONTENTS

MULTIMODAL CHOICE MODELLING - SOME RELEVANT ISSUES

1. INTRODUCTION

The problems of mode choice modelling and forecasting have been approached in many ways since the mid-50s, but for the most part, research and applications have been concerned with choice between car and public transport which, it has been argued, is the situation faced by the majority of travellers in the journey-to-work. However, it is obvious that people do not necessarily choose between two specific alternatives only when making their choice, but instead they generally confront options such as driving a car, travelling as passengers in a car, bus or train, riding a bicycle or a scooter or simply walking. In addition, each trip might utilise a combination of modes, i.e. mixed-mode trips (for example, park-and-ride), although it can be argued that some mixed options are so unlikely that the probability of their selection can be considered as zero. As a consequence, it has often been suggested that individuals can be considered as users of their 'main mode' (e.g. the procedure used in the majority of transportation studies in the U.K.). However, this procedure is clearly inaccurate for many people who depend on another mode for access to the major one. Also, with the increasing departure from traditional policies based on a 'pure' mode context and the emphasis on an 'integrated' approach to urban transport problems, the time is ripe for models which are more oriented towards alternative policies, such as price penalty measures, traffic restraint and exclusion schemes, bus priority measures, incentives to park-and-ride and car-pooling, etc., and which must be multimodal (as opposed to binary) in nature.

During the last decade, and particularly over the last five years, significant advances have been made in travel demand forecasting methods. The most important and widely promoted new techniques have been the so-called 'disaggregate' or 'individual-choice' or 'second generation' models (for a good review of theoretical developments, see Williams, 1979). These models have been usually generated within a 'random utility' theory framework(*) (for a review, see Domencich and

... ...

(*) Note that the theory is not constrained to disaggregate models only; in fact we have used it recently to generate aggregate modal split models (see Hartley and Ortuzar, 1980).

McFadden, 1975). In this quantal choice theory, the decision-maker is assumed to choose the option (A_j) which possesses, as far as he is concerned, the greatest utility U_j. In order to account for dispersion - the fact that individuals with the same observable characteristics do not necessarily select the same option - the modeller introduces a random element ε_j in addition to each measured individual's utility $\overline{U}_j$. In this way, the utility of alternative A_j is actually represented as:

$$U_j = \overline{U}_j + \varepsilon_j$$

Ortuzar and Williams (1978) have described pedagogically, the generation of random utility models, ranging from the very convenient but theoretically restrictive multinomial logit (MNL) model, to the general and powerful but rather intractable multinomial probit (MNP) model.

In this paper we wish to discuss briefly some issues related to the application of such models (and in some cases any model) to the problem of choice of mode for the journey-to-work. We will consider questions of data, such as type of data, alternative sampling strategies and problems of measurement, and modelling issues, such as model specification and estimation. However, we will first address the aggregation problem which lies at the heart of one of today's most hotly contested debates - whether to use aggregate or disaggregate models, and in which circumstances.

We do not attempt to be comprehensive on these issues, so we refer the reader to good general discussions by McFadden (1976; 1979a); Williams (1977; 1979); Hensher (1979a); Ben-Akiva et al (1977; 1979); Daganzo (1980); Daly (1979); Jansen et al (1979); Manheim (1979); Reid (1977); Spear (1977; 1979); and Williams and Ortuzar (1980b).

2. THE PROBLEM OF AGGREGATION

The aggregation issue may be thought of in very general terms as the path through which a detailed description of an individual's decision-making process, as imputed by a modeller, is transformed into a set of observable entities and for relations which can be usefully employed by him. In an econometric interpretation of (transport demand) models, the aggregation over *unobservable entities* results in a probabilistic decision (choice) micro model, and the aggregation over the *distribution of observable quantities* results in the conventional

aggregate or macro relations. In this sense, the difficulty of the aggregation problem depends, to a large extent, on how the components of a system are described within the frame of reference used by a modeller, because it is precisely this framework which will determine the degree of variability to be accounted for in a 'causal' relation.(*) To give an example, if the frame of reference used by a modeller is, say, that provided by the entropy maximising approach, the explanation of the statistical dispersion in a given data set will be very different to that provided by another observer using a random utility maximising approach, even if they both finish up with *identical model functions* (the equi-finality issue, see, for example, Williams, 1979). The interpretation of such a model, say the classicial MNL, depends however on the theory used to generate it, and this is particularly important for its elasticity parameters. For the entropy maximising modeller, the parameter corresponds to a Lagrange multiplier associated

> "..... *with the change in likelihood of observing a given allocation (share) pattern ... with respect to incremental changes in system trip cost measures". (Williams, 1979)*

For the second modeller, the same parameter is now inversely related to the standard deviation of the utility distributions from which the model is generated(**) see Williams (1977).

If we choose to use a random utility approach, the aggregation problem will reduce, to obtain from data, at the level of the individual, aggregate measures such as market shares of different modes, flows on links, etc., which are typical final model outputs. There are two obvious ways of proceeding, as shown in Figure 1(a), which are basically distinguished by having the process of aggregating individual data *before* or *after* model estimation. If the data is grouped prior to the estimation of the model, we will have the classical 'aggregate' approach which has been heavily criticised for being inefficient in the use of data (because data is aggregated, each observation is not used as a data point and therefore more data is needed), for not accounting for

... ...

(*) I am grateful to Huw Williams for having explained this interpretation to me.

(**) Two comments are worthwhil e here: firstly the full interpretation of model parameters is not transferable within theories; and, secondly, while in some cases the interpretation might not matter (i.e. if one is interested on flows in networks) in others it can be very crucial, for example, if we are seeking to endow predicted forecasts with some notion of benefits (Williams and Ortuzar, 1980b).

the full variability in the data (e.g. within zone variance may be higher than iner-zonal variance), and for risking statistical distortion and bias (such as the well-known ecological fallacy), etc. The 'disaggregate' approach, on the other hand, estimates the model at the level of the individual thus apparently answering, at this stage, the criticisms mentioned above. The question that remains, however, is how to perform the aggregation operation over the micro relations? As we will see below, the answer is ... 'rather simply', *if* we are interested in short-term predictions of journey-to-work mode choice models; however, for other modelling requirements, the answer ranges from ... 'difficult', to ... 'almost impossible', unless being self-defeating in the sense of requiring heroic assumptions (as bad as those criticised in the 'aggregate' approach) and/or enormous amounts of extra data. In fact, Reid (1977) in the context of developing a disaggregate model system has remarked that

> "... there are practical and theoretical limits to the application of strictly behavioural methods ... it is difficult to preserve a behavioural structure and conform to aggregate observations..."

Before briefly describing the main aggregation methods, let us note that the approach followed in British practice is a hybrid of the two mentioned above as shown in Figure 1(b). For example, household based (rather than zonal) category analysis has been used at the trip generation stage, while the SELNEC and subsequent studies used weighting coefficients obtained from a standard disaggregate study (e.g. McIntosh and Quarmby, 1970), in a generalised cost formulation. However, the elasticity parameters (e.g. β and λ) and other model constants have been determined from an aggregate calibration. This 'transferability' of micro parameters (*) between different studies (e.g. different regions and different times) with the possibility of local 'tuning' (Goodwin, 1978) may be seen as a pragmatic approach to the aggregation problem. This issue is discussed at more length by Williams and Ortuzar (1980b).

... ...

(*) Which interestingly bears close analogy to the strategy proposed by Ben-Akiva (1979) for the transferability of disaggregate models, although with different motivations.

Returning to the general approaches shown in Figure 1a, much research has been directed recently at a comparative assessment of aggregation methods (see, for example, Ben-Akiva and Atherton, 1977; Ben-Akiva and Koppelman, 1974; Bouthelier and Daganzo, 1979; Daly, 1976; Dehghani and Talvitie, 1979; Hasan, 1977; Koppelman, 1974, 1976a, 1976b; Liou et al, 1975; McFadden and Reid, 1975; Meyburg and Stopher, 1975; Miller, 1974; Reid, 1978a, 1978b; Ruijgrok, 1979; Watanatada and Ben-Akiva, 1978). The various methods proposed offer different strategies for computing the summation/integration over micro relations, and include, among others: the 'naive' approach, sample enumeration methods, and classification approaches.

The naive approach consists of the direct substitution of aggregate or average values of the explanatory variables into typically non-linear, micro relations, and it has been found that the aggregation bias may be severe in this case. In the sample enumeration approach, the impact of a given policy on each individual, in a representative sample, is determined from the disaggregate model and population forecasts are then computed by straightforward summation of the effect over individuals according to the sampling strategy. This method is considered to be particularly useful for estimating impacts for *short-term* policies (see Ben-Akiva and Atherton, 1977), but must be modified when the characteristics of the population change over the forecasting period (since it cannot be assumed that the distribution of observable attributes remains constant).

In the classification approach, the total population is partitioned into relatively homogeneous groups and then average (group) values of the explanatory variables are inserted into the disaggregate model to determine demand in each group according to the naive approach(*). The accuracy and efficiency of the method depends on the classification involved, e.g. the type and number of groups and the characteristics of the variables included.

... ...

(*) In terms of its aggregation characteristics, the practice in British studies with use of market segment differentiated models, may perhaps best be seen as a variation of this classification approach.

For *mode choice* studies where only short term elasticities are required, there is a consensus that aggregating micro-relations, i.e. the 'disaggregate' approach, is both feasible, efficient and hence desirable. However, in longer term contexts where location (distribution) models need to be considered and/or when network flows are required the problem becomes much more involved. Very few studies have attempted the aggregation of micro-models in these contexts so it is premature to make definitive judgements. One which did, the SIGMO study (Project Bureau Integral Traffic and Transportation Studies, 1977) encountered severe problems in attempting to reconcile micro destination choice models with aggregate trip patterns and abandoned the disaggregate approach in favour of an existing distribution model based on generalised costs. More generally, Reid (1977) has noted that while in principle a disaggregate model has a better chance of capturing the essential causality in the data, in practice

> "... if the behavioural theory is weak or the models untested against experience, such as with current individual location models, they may fail to include some important factors which are embodied in aggregate or summary variables which merely show a correlation to demand. These are more likely to pick up unknown effects ... (and) ... if adequate disaggregate data will not be available for forecasting, models calibrated on aggregate data will be more accurate."

In the early 1970's the process of aggregation was usually viewed as the rather straightforward solution of a numerical problem which was well understood in principle. In practice, however, it has shown itself to be a highly non-trivial process which embraces not only considerations of numerical efficiency, but also questions relating to the availability of forecasts for individual explanatory variables and the stability of the distribution of explanatory variables over time. Furthermore, there is also concern about the relation of predictions to estimation and data designs; therefore, any comparison of 'aggregate' and 'disaggregate' models must involve, implicitly or explicitly, a consideration of these issues.

3. DATA COLLECTION AND MEASUREMENT

3.1 Representation and measurement of travel attributes

In any particular study, out of the large variety of potentially available forecasting methods (e.g. cross-sectional analysis; panel data methods; aggregate time series approaches) and estimation techniques, data considerations alone will normally restrict the choice to one single method. Historically, the cross-sectional approach has clearly dominated, typically in conjunction with revealed preference methods, although alternative approaches based, for example, on stated preferences/intentions, have been preferred on several occasions (see Ortuzar, 1980a). However, the general problem of discounting for the over-enthusiasm of respondents (the 'yeah' bias) has not yet been solved, and it has recently been suggested that stated and revealed preference methods may perhaps be better used in a complementary fashion, where insights can be obtained which would not arise if either approach were used alone (see, for example, Hensher and Louviere, 1979; Gensch, 1980). We have argued elsewhere, (Williams and Ortuzar, 1980a), that it is not possible at the cross-section to discriminate between a large variety of possible sources of dispersion in data patterns (such as preference dispersion, constraints, habit effects, etc.). Panel data or more simply, before-and-after information, may offer some means to directly test and perhaps reject hypotheses relating to response, (see an interesting example in Johnson and Hensher, 1980). On the other hand models built on 'longitudinal' (as opposed to cross-sectional data) have technical problems of their own (e.g. how best to 'pool' the information), but a discussion of their merits is beyond the scope of this paper.

A related area of concern has to do with the problem of measurement. We wish to discuss briefly here the implications for parameter estimates of using different measurement techniques and/or philosophies. For a deeper insight into the problem we refer the reader to the excellent discussions by Daly (1978) and Bruzelius (1979). The problems involved in obtaining measures of explanatory variables (e.g. cost and time requirements by alternative modes) are shown schematically in Figure 2. Ideally we would like to obtain information on these variables as *perceived* by the commuter when taking his decision, especially if we are *not* interested in forecasting (how do you get 'perceived' information

about a future situation?), but perhaps in obtaining 'values of time'. The figure reflects the state-of-the-art in the understanding of the relationships between 'actual', 'perceived', 'reported' and 'measured' values. The trouble is that none of the arrows and boxes in the figure have yet been quantified. Knowledge in this area is, literally, sketchy! The analyst is therefore made to choose between reported and measured (or 'engineering' or 'synthesised') data, and while models estimated on each type of data may prove reasonable in themselves

> "... it is very difficult to postulate relationships that will allow models calibrated on reported data to be applied to synthesised data or vice versa." (Daly, 1978)

Most probably the safest way out is to collect information on both reported and engineering values and to make comparisons in order to gain insight from the two approaches. This is, of course, more costly and time consuming and, as Hensher (1979c) and others have remarked, it is seldom the case that the analyst finds himself with the luxury (or embarassment) of alternative data/methods at hand.

We mentioned above that one possible and alternative use for a model, instead of forecasting, is to employ it for estimating, for example, values of time (Bruzelius, 1979; Daly, 1978; Hensher, 1972; McFadden, 1978b; Prashker, 1979; Quarmby, 1967; Train, 1977; Gunn, Mackie and Ortuzar, 1980; and some of the references cited therein). An old issue in this context is the 'trader/non-trader' question, e.g. should those individuals who appear to be faced with a dominant(*) option be excluded from the sample? As Daly (1978) has clearly pointed out, the answer is definitely no! The main difficulty has actually been due to a misunderstanding: that only *observable*, and hence measured (or measurable) attributes should matter when defining whether an option is dominant, leaving out the crucial *unobservables* and/or unmeasured characteristics. In this sense, the larger the number of measured attributes incorporated in the model, the smaller will be the number of apparent 'non-traders' and, better still, the less the influence of unmeasured factors (simply because more of those are incorporated.)

... ...

(*) An option which, *to the modeller*, looks better in every respect than the others and happens to be the chosen one (if it is not the chosen one the individual is deemed irrational!). Notice that this is not to be confused with the issue of *captive* travellers (e.g. a person who needs the car during the day) who should be trimmed out of the sample (if identified).

This brings us naturally into the question of using attitudinal variables (eg. comfort, convenience, reliability) to improve our models. (For a more complete discussion see, Foerster, 1979b, Johnson, 1975; Spear, 1976; Stopher *et al.* 1974; and Wermuth, 1978). In terms of the influence of attitudinal measures on the value of other parameters and on the general performance of a model, there is conflicting evidence in the literature. McFadden (1976), for example, concluded that choice was explained, to a great extent, by the typical level-of-service variables used in conventional studies and that attitudinal measures added very little explanatory power to the models(†). More recently, however, Prashker (1979) has found that including measures of reliability (eg. reliability of finding a parking space; reliability of bus arrivals), both substantially increased the explanatory power of the models (for example, it produced mode-specific constants which were not statistically different from zero), and change significantly the values of some parameters (in particular the value of in-vehicle time). Once more, the safest recommendation seems to be to examine the possibility of measuring some 'unconventional' factors (eg. reliability, comfort, convenience, etc.) and to test for their effects on the other parameter estimates and model explanatory power. Again, however, this would naturally imply higher data collection and analysis costs.

... ...

(†) It is fair to say, though, that the models discussed by McFadden (1976) have been heavily criticised by Talvitie and Kirshner (1978) on the grounds, among other things, that the mode-specific constants tended to account for over 60% of their explanatory power.

3.2 Alternative sampling strategies

The development and implementation of travel demand models have traditionally been associated with large data collection efforts, involving, principally, very expensive home interview surveys. Because conventional aggregate models used data at the zonal level fairly large random samples were required for calibration purposes, and it is well-known that on many occasions the cost and time consumed in the collection and analysis of the data prevented the analysts from examining a sufficient range of alternative policies.

One of the advantages traditionally cited for disaggregate models is the efficiency with which they can make use of available data and the potential for reducing the time and effort expended on data collection. As we saw above, this claim (together with others) has not been universally achieved, but it is true to say that in certain situations the fact that disaggregate choice models use observations of individual decision makers, rather than geographically defined groups, can substantially reduce data collection costs. The rest of this section summarises two excellent papers by Lerman and Manski (1976; 1979) which constitute the state-of-the-art in this area.

The majority of applications of disaggregate choice models have relied on randomly sampled data, eg. slight variations on the typical home interview survey. A few studies have used stratified sampling, where the population of interest is divided into groups according to some characteristics such as car ownership (which must be known in advance) and each subpopulation is sampled randomly. It is clear that random or stratified samples can be

very expensive indeed in cases where an option of interest has a very low probability of selection; because to achieve a reasonable representation of the option in question it is necessary to collect a very large sample. A choice-based sample (that is, one where observations are drawn based on the outcome of the decision-making process under study) designed so that the number of users of the low option is predetermined offers one way to solve this problem.

Choice-based samples are not uncommon in transport studies. Typical examples are on-board train and bus surveys, and roadside interviews in the case of mode choice modelling. They can frequently be obtained fairly inexpensively, but (because of the way the parameters of (disaggregate) models are generally calibrated) have seldom been used for calibrating models (see Cosslett, 1980). As we will see below each sampling strategy results in a different distribution of observed choices and characteristics in the sample that in certain situations the fact that disaggregate choice models use observations of individual decision makers, rather than geographically defined groups, can substantially reduce data collection costs. The rest of this section summarises two excellent papers by Lerman and Manski (1976; 1979) which constitute the state-of-the-art in this area.

The majority of applications of disaggregate choice models have relied on randomly sampled data, eg. slight variations on the typical home interview survey. A few studies have used stratified sampling, where the population of interest is divided into groups according to some characteristics such as car ownership (which must be known in advance) and each subpopulation is sampled randomly. It is clear that random or stratified samples can be

very expensive indeed in cases where an option of interest has a very low probability of selection; because to achieve a reasonable representation of the option in question it is necessary to collect a very large sample. A choice-based sample (that is, one where observations are drawn based on the outcome of the decision-making process under study) designed so that the number of users of the low option is predetermined offers one way to solve this problem.

Choice-based samples are not uncommon in transport studies. Typical examples are on-board train and bus surveys, and roadside interviews in the case of mode choice modelling. They can frequently be obtained fairly inexpensively, but (because of the way the parameters of (disaggregate) models are generally calibrated) have seldom been used for calibrating models (see Cosslett, 1980). As we will see below each sampling strategy results in a different distribution of observed choices and characteristics in the sample and hence each has associated a different calibration function (such as likelihood). Although the first two sampling methods present no problems to existing software, the choice-based approach needs some modifications (Lerman, Manski and Atherton, 1976; Lerman and Manski, 1976) or existing programs will produce biased parameters(*).

Given the existence of a practical estimation procedure for choice-based samples, the question is what sampling strategy should be preferred. Lerman and Manski (1976; 1979) have argued that unfortunately, the answer is extremely situation-specific and depends on

... ...

(*) For a practical application (if rather a 'pragmatic' one) of the use of existing software to estimate disaggregate models from a choice based sample refer to Stopher and Wilmot (1979).

- the cost of various sampling methods
- the choice being modelled
- the characteristics of the population under study
- the social cost of estimation errors in terms of applications of misguided policies(†)

Random samples often require a major expenditure of time and money to collect. Normally they should be based on homes - if done anywhere else they would be choice-based because the respondent has already made a trip choice - with all the problems associated with home interview surveys. However there is scope for longer and more in-depth interviewing.

A further problem of random samples is that they offer no opportunity to increase the amount of information given a fixed sample size. Variation in the data(*) cannot be controlled in this case, being rather a random outcome of the sampling process. Stratified samples on the other hand should help in this sense, because even if the characteristics of the population vary little, the sample itself can have a high variance, ie. certain strata can be sampled at different rates from others. However, stratified samples are often more expensive than random ones because, in order to sample at random from a subpopulation, one must first be able to isolate the subpopulation; in practice this may be difficult (and expensive) to achieve(**).

... ...

(†) See Gensch (1980) for an interesting example about the possible magnitude of such costs.

(*) The more variation in the data, the more reliable are the parameter estimates.

(**) For example one may need to begin an interview to find out the stratum to which the respondent belongs.

In general choice-based samples are the least expensive but they require prior knowledge of the ratio of the share of the entire population choosing each alternative to the sample share. Fortunately, the former is an aggregate statistic which might be obtained from several sources (Lerman and Manski, 1976). Another problem of this sampling strategy is that of bias (*), or alternatively, how to ensure that the sample, given the users of an option, is random. Lerman and Manski (1979) mention as an example the problem, in an on-bus survey, of allowing for the fact that some routes may have a higher percentage of elderly users while others may attract primarily workers. Another case is that associated with high rejection rates of mail-back questionnaires where it is unlikely that the distribution of characteristics of those who choose to respond will be the same as that of the population as a whole.

Bearing all the above issues in mind, Lerman and Manski (1976) concluded in their paper

> "... In all probability the question of sample design will remain a judgemental problem."

and we see no reason why we should challenge this view.

4. Model Specification

Having available, or having decided to collect data in a certain way and of a given type - typically a random sample of cross-sectional information on revealed preferences, where values of attributes are either measured or synthesised - the analyst still has some options open in terms of the model structure,

... ...

(*) A problem of stratified samples in general.

specification and estimation method to use. In section 5 we will present a fairly comprehensive review of the most widely recommended method of estimating discrete choice models - Maximum Likelihood (ML) estimation - with particular emphasis on disaggregate data. (Elsewhere, (Hartley and Ortuzar, 1980), we have discussed the method as applied to the calibration of aggregate hierarchical logit modal split models and compared it with alternative procedures.) Firstly though, we wish to briefly comment here on the related problem of model selection in general.

4.1 Model selection

In general, the structure of a model, the variables entering it and their form, the form of the utility functions themselves, and so on, are all matters for testing and experimentation (see the excellent book by Leamer, 1978), and are quite often a strong function of context and data availability. Aggregate models have often been critically viewed as policy insensitive, either because a key variable has been completely left out of the model; or from some component(s) of the model thought to be sensitive to it (eg. inelastic trip generation); or because severe distortions could be introduced from specification or aggregation bias errors. In this sense the American UTPS system was particularly weak (Ben-Akiva *et al.*, 1977).

In British practice, however, the concept of generalised costs, together with network modifications, have been used to test a very wide range of policies (eg. from road investments to parking restraint and park-and-ride systems), although these have only been interpreted on terms of the variables(*): in-vehicle-time, out-of-

... ...

(*) Although disaggregate models include many more explanatory variables, including socio-economic, level-of-service and even attitudinal variables, we mentioned in section 3 that most of the statistical explanatory power of the models (excepting the large amount explained by mode-specific constants, Talvitie and Kirshner, 1978) rests in relatively few of these attributes, including the usual level-of-service variables (McFadden, 1976).

vehicle time and out-of-pocket costs (suitable scaled by the generalised cost coefficient). Also a large variety of model structures have been employed (see the discussion by Williams, 1979) including both simultaneous and sequential model forms, and the policy responsiveness of models has been found to be critically dependent on model specification, to the extent that certain models since have been recognised as 'pathological' (i.e. implied elasticities of the wrong sign) because their structures were not properly diagnosed for specification errors (see Senior and Williams, 1977; and Williams and Senior, 1977).

The consideration of available alternatives (which could also be discussed as an aggregation issue) is another part of the specification process with strong implications for policy sensitivity. In the vast majority of aggregate studies only binary choice between car and public transport has been considered, with the consequence that the multimodal problem has not been treated very seriously. In the best cases the consideration of alternative public transport options has been relegated to the assignment stage, employing 'all-or-nothing' or 'multipath' allocation of trips to sub-modal network links. We have given elsewhere, (Hartly and Ortuzar, 1980), a practical example of fitting a rather more general structure than the simple MNL to aggregate modal split data for three modes (car, bus and train) and show how a priori notions which led us to postulate such structure were confirmed by appropriate structural diagnosis tests. Here we will concentrate on disaggregate models both because the full range of issues in their specification are more apparent and because they have been more thoroughly aired and discussed.

We mentioned above that the final specification of a model tends to be a strong function of context and data availability. A priori notions and theoretical insight also provide valuable help while another important pragmatic factor is the availability of specialised software. In fact, one reason why linear-in-the-parameters logit (and simple binary probit) models have been so popular is that they can easily be estimated using available software (for well documented examples, see Boyce, Desfor, et al., 1974; Domencich and McFadden, 1975; Ben-Akiva and Atherton, 1977; Hensher, 1979c; and Talvitie and Kirshner, 1978) whilst other more general forms normally present enormous difficulties (see the discussion on probit models by Sheffi, Hall and Daganzo, 1980).

On the other hand, the limitations of 'simple scaleable choice models' typified by the MNL structure have been one of the prime motivations behind the interest in alternative models of the decision process; although we have argued elsewhere (Williams and Ortuzar,1980a) that, in a certain sense, the development of more general random utility structures (such as the MNP) has removed some of the original justifications for building such models. However, this does not mean that the more conventional models are necessarily appropriate; indeed, it is often useful and desirable to examine competing frameworks. One cause for concern, though, is that different model structures and forms tend to produce different parameter estimates and response elasticities, whilst we do not have means to discriminate between them at the cross-section (see Williams and Ortuzar, 1980a).

4.2 Choice set determination

One of the first problems an analyst has to solve, given a typical (i.e. as defined above) data set is that of deciding which alternatives are available to each individual in the sample. As Hensher (1979c) has noted

> "... Choice set determination ... is the most difficult of all the issues to resolve. It reflects ... the dilemma which a modeller has to tackle in arriving at a suitable trade-off between modelling relevance and modelling complexity. Usually, however, *data availability acts as a yardstick.*" (our emphasis)

It is extremely difficult to decide on an individual's choice set unless one asks him; therefore the problem is closely connected with the already discussed dilemma of whether to use reported or measured data. The obvious procedures of (a) taking into account only those alternatives which are effectively chosen in the sample; or (b) to assume that everybody has all alternatives available (and hence let the model decide that the choice probabilities of the unrealistic alternatives are low or zero) have also obvious disadvantages. For example, in the former case it is possible to miss realistic alternatives which are not chosen

(due to the specific sample or sampling technique). In the latter case, the inclusion of too many alternatives may affect the discriminatory capacities of the model, in the sense that a model capable of dealing with unrealistic alternatives may not be able to describe adequately the choices among realistic options (see, Ruijgrok, 1979). Fortunately, in the context that interest us here - mode choice modelling - the number of alternatives is usually small and the problem should not be severe.

By contrast, in destination choice modelling (ie. trip distribution) the identification of alternatives in the choice set is a crucial matter, and not simply because the total number of alternatives is usually very high(*). To illustrate this, consider the case of modelling the behaviour of a group of individuals who vary a great deal in terms of their knowledge of potential destinations (owing perhaps to varying lengths of residence in the describe the relationship between predicted utilities and observed choices, may be influenced as much by variation in choice sets among individuals (which are *not* fully accounted for in the model) as by variations in actual preferences (which are accounted for). Because changes in the nature of destinations may affect both choice set and preferences to different degrees, this confusion may be likely to play havoc with the use of the models in forecasting or in the possibility of transferring their specification over space. It is interesting to note in this context that McFadden (1978a) has shown that for a MNL, the model parameters can be estimated without bias by sampling alternatives at random from the full set of options, with appropriate adjustments in the estimation mechanisms. This is,however, not possible for the

... ...

(*) Although this in itself is also quite a problem because current software is only capable of dealing with 20 to 30 options.

MNP, for example, precisely due to its improved specification which allows for interaction between all alternatives.

4.3 Defining the form of the utility function

Another area of concern in 'specification searches' relates to the form of the utility functions. Although there is broad agreement among experts that for mode choice modelling the *convenient* assumption of 'representative' utilities with linear-in-the-parameters (LTP) forms should present little difficulty, in other contexts such as destination choice modelling(*) the general agreement is that LTP utility functions are not valid (see, for example, Foerster, 1979a; Daly, 1979; Louviere and Meyer, 1979). The problem this time is partly the lack of appropriate estimation software, and partly theoretical(**). Three general approaches have been proposed to deal with this problem:

- the use of functional measurement/conjoint analysis techniques with experimental design data (Lerman and Louviere, 1978; Hensher, 1979a, 1979b; Hensher and Louviere, 1979).
- the use of 'form searches' by means of statistical transformations (e.g. the Box-Cox method) as in the work of Gaudry and Wills (1977).
- the constructive use of the economic theory itself for the derivation of form (Train and McFadden, 1978; Hensher and Johnson, 1980).

Exploring this issue further would be outside the scope of this paper but we wish to mention not only that non-linear utility forms imply different trade-off mechanisms than those usually associated with a concept like the 'value-of-time'; but also, and more importantly, that model elasticities and forecasting power have been shown to vary dramatically with functional form (see, Dagenais, Gaudry and Liem, 1980). Thus the issue has important implications for model design and hypothesis testing.

... ...

(*) A further major challenge in destination choice modelling (and in addition in mode choice modelling for non-work journeys such as shopping trips) is how to measure and/or represent the attractiveness of destinations. For the case of mode choice for the journey-to-eork this is not a problem because in the short term it can be safely assumed that destinations are fixed; therefore, their attractions are common to all competing modes and thus cancel out. When this assumption does not hold (as is the case with shopping trips) we face a problem which has, so far as we are aware, no satisfactory answers.

(**) Specifically the problem is that for non-linear utility expressions there is no guarantee that the likelihood function has a unique optimum (Daganzo, 1979).

4.4 Model structure and variable selection

Having solved or simply avoided (as in our case) the aforementioned problems we have to deal with two further obstacles:

- what model form (and structure) to use, eg. logit or probit
- given the structure, *what* variables should enter the utility functions and in what *form*

We think it is fair to say that the question of model structure can only be resolved by examining the particular situation under study. If we have reasons to believe that alternatives are independent and that variations in taste among individuals in the population are not important (e.g. we can speak of a single value, rather than a distribution, for the coefficients multiplying the attributes entering the utility functions), then we may confidently choose the MNL model. If, on the other hand, the above conditions are not met or if we are not certain, then we *should* test alternative (more complex) model structures against the convenient MNL. For example, if we suspect that correlation between alternatives may be a serious problem, we can either test if the 'independence from irrelevant alternaties' condition is satisfied (McFadden, Tye and Train, 1976) or, better still, estimate a hierarchical logit model which includes built-in structural diagnosis tests (Sobel, 1980; Ortuzar, 1980b; Ortuzar 1980c). On the other hand, if we have reasons to believe that there are strong taste variations effects, we might have to try and fit a 'random coefficients' model. The simplest one is the CRA Hedonics model (Cardell and Reddy, 1977) which still has the restriction of assuming non-correlated alternatives as the MNL. The most general model structure possible, and sadly the more complex to estimate(*), is the MNP model which allows for the existence of both correlation and taste variations in the data.

It is important to realise that use of an inadequate model, such as the MNL, can lead to serious errors (Hausman and Wise, 1978; Horowitz, 1978, 1979a, 1979b, 1980) and studies on the comparison of alternative

... ...

(*) The special problems of estimating probit models are discussed by Sheffi, Hall and Daganno 1980. The interested reader is also referred to the excellent book by Daganzo (1980).

model structures using simulated data, such as those described in Ortuzar (1978, 1979, 1980a) and Williams and Ortuzar (1980a) among others, have tended to confirm this view.

Even if the analyst is convinced (or has no choice but to be convinced) that a given model structure (say a MNL model) is adequate and that linear-in-the-parameters utility functions pose no difficulties, he has still to decide what variables should enter the utility expressions, and in what form. This question is particularly relevant in the case of socio-economic variables. In disaggregate modelling work the most common approach until the mid-1970's was to add these variables as additional linear terms; this is consistent with the hypothesis that any trade-off mechanisms involving say, time and costs, are the same for all individuals.

Two alternative approaches allow different trade-off functions for groups of people with different characteristics. The first, which is fully consistent with the requirement of observing groups of individuals with the same choices and constraints, is to stratify the sample on the basis of the individual characteristics and to calibrate a model for each market segment. In this way the model coefficients are allowed to vary for the different market segments, thus resulting in potentially different trade-off mechanisms(*). The problem is, as usual, one of data: the larger the number of market segments, the smaller the number of observations on each for a given sample size. The second one, which can be used in conjunction with the first, is to express certain coefficients (eg. of the time or cost variables) as a function of an individual descriptor, usually income (see the

(*) This is not to be confused with the issue if random vs. fixed coefficients models as discussed above. Here we are simply considering fixed coefficient models being applied to different market segments.

discussion by Train and McFadden, 1978). In a value-of-time context this would, for example, result in time being valued as a percentage of the wage rate (McFadden, 1976).

The decision about what variables enter the utility function and in what form (eg. level-of-service variables being generic or mode-specific, etc.) is usually approached in a stepwise fashion by testing if the extra variable or form adds extra explanatory power to the model. This is related to questions of model credibility and policy sensitivity in the following sense; it may often occur that a variable which is considered to be important, either on strong a priori grounds or because it is a key one in the policy-model interface (eg. a cost variable in a study of pricing mechanisms), would be left out as statistically insignificant by a stepwise selection procedure. In such a case, the tendency has been to override the 'automatic' selection procedure (see Gunn and Bates, 1980). The stepwise selection of variables is usually done as part of the model estimation phase; so we will postpone a discussion on methods to do this until section 5.2.

5. MODEL ESTIMATION

5.1 General statement of the problem(*)

In travel demand modelling (as in most modelling exercises) interest centres on finding a *causal* relationship between one variable, or set of variables, held to be dependent on another variable, or set of variables. The purpose of the exercise is to predict what value the dependent variable will take given particular known or hypothesised (forecast) values of the

(*) I will draw heavily here on unpublished seminar notes by Hugh Gunn, with whom I have also benefited greatly from discussions in all aspects relating to the statistical interpretation of models.

explanatory variables. For two variables we can simply write

$$Y = f(X) \tag{1}$$

and the problem is approached by collecting a sample of, say, n pairs of observations $\{x_i, y_i\}$, i=1, ..., n, and letting the data determine the 'best' form of $f(\cdot)$. On certain occasions, given enough data points, no mathematical analysis is needed; for any given (forecast) value of, say, x_o, we simply consult the data, find the nearest observed value of x to x_o and use the corresponding value y as the modelled result. With less data we will normally need to interpolate values, or at a considerably greater risk, extrapolate them. For this we need to assume a functional form for $f(\cdot)$; an estimation problem arises when the relationship between Y and X is not exact. Formally, we can postulate the model form:

$$y_i = f(x_i) + \varepsilon_i \tag{2}$$

where the error term, ε_i, is introduced to account for the scatter in the data. Estimation consists of choosing particular values for the unknown coefficients in f(X) in order to minimise the 'distance' between modelled and observed values of the dependent variable at the set of data points. In other words we want to maximise the similarity between Y and f(X) and for this we must choose a suitable measure of 'distance' from the many available, such as

$$D_1 = |Y - f(X)| \tag{3}$$

$$D_2 = (Y - f(X))^2 \tag{4}$$

$$D_3 = \left(\frac{Y - f(X)}{Y}\right)^n \tag{5}$$

etc.

Each criterion of goodness-of-fit will determine a corresponding set of estimates of the unknown coefficients - the problem is which is the 'best'(*)?

When the error terms are each independent with mean zero and constant variance, D_2, the least squares criterion, is known to give such 'best' estimates *on average*(**). For general error distributions, which may vary from observation to observation, a satisfactory criterion of fit must allow for the relative reliability of each data point. The method of Maximum Likelihood (ML), which we will describe below, does just that and it is interesting to note in passing that if the errors ε_i have common and independent Normal distributions, the criteria ML and D_2 are identical. For models in which the dependent variable Y is a proportion (such as in the case of an aggregate modal split model), it appears sensible to choose f(X), such that

$$0 < f(x_i) < 1 , \quad \forall x_i \varepsilon X \qquad (6)$$

If we move to a more general case where we wish to model an exhaustive set of outcomes $\{Y^1 Y^2 \ldots Y^N\}$ where

$$\sum_{j=1}^{N} Y^j = 1 \qquad (7)$$

then it is also sensible to ensure that the models $\{f_1(X^i) f_2(X^i) \ldots f_N(X^i)\}$ are such that

$$\sum_{j=1}^{N} f_j(X^i) = 1 \qquad (8)$$

... ...

(*) Usually interpreted as the most reliable in terms of the forecasts it produces.

(**) However, problems arise with its use when different data points have different errors - weights may have to be introduced, or transformations made (see, for example, Bishop, Fienberg and Holland, 1975).

Now if we choose

$$f_j(X^i) = K_i \exp\{g_j(X^i)\} \qquad (9)$$

we will ensure that the non-negativity condition in (6) is satisfied for any function $g_j(\cdot)$, if K_i is a positive constant.

Furthermore setting

$$K_i = \frac{1}{\sum_{j=1}^{N} \exp\{g_j(X^i)\}} \qquad (10)$$

ensures that the models sum to unity. The combination of (9) and (10) is, of course, the logit model which has been used for decades to analyse tables of proportions for precisely the reasons given above. Thus the random utility generation of the model has been a post-hoc rationalisation for use of the model in certain circumstances where it might be appropriate (for a fuller discussion of this issue, see Williams and Ortuzar, 1980b).

5.2 Maximum Likelihood (ML) estimation and allied statistical tests

ML calibration of aggregate nested logit models (as series of logit models), together with a discussion of ML and other calibration methods for aggregate data (eg. where proportions rather than (0,1) choices are observed) have been presented in Hartley and Ortuzar (1980). Here we will concentrate on the special problems arising in the estimation of any disaggregate model. The differences stem from the basic fact that while models predict choice probabilities (i.e. numbers between 0 and 1), they must be tested and calibrated against (0,1) choice behaviour (*). From now on, we will assume

... ...

(*) For a good general discussion of the problems involved, the reader is referred to McFadden (1976); Stopher (1975); Tardiff (1976); Hauser (1978) and Project Bureau Integral Traffic and Transportation Studies (1977).

that the modeller has gathered, following a certain sampling rule, information on the actual choices (eg. alternative A_i, from the choice set $\underline{A}(q)\varepsilon\underline{A}$) of individuals q, and information on choice influencing variables Z_{iq}^k (these may be level-of-service attributes of the options and/or socio-economic characteristics). The ML technique, which has been the most widely used and more strongly recommended method (Jansen *et al.*, 1977; McFadden, 1976, 1979b) looks at the probability of obtaining the Q independent choices, C_q, q=1, ..., Q, given the model (along with its parameters $\underline{\Theta}$):

$$P(c_q,\underline{\Theta}) \qquad (11)$$

Then the probability of obtaining the observations $c_1, c_2, \ldots, c_Q$, is

$$L(c_1, \ldots, c_Q,\underline{\Theta}) = \prod_{q=1}^{Q} P(c_q,\underline{\Theta}) \qquad (12)$$

The usual way of looking at this function is to regard the vector of parameters $\underline{\Theta}$ as known and L as a set of probabilities over possible observations. However, in the estimation context, the observations are known and $\underline{\Theta}$ is unknown. When L is regarded as a function of $\underline{\Theta}$ for given (observed) c_q, q=1, ..., Q, it is called the 'Likelihood Function' and is normally written as $L(\underline{\Theta})$, for short. Recall that the observed dependent variable takes a value of either 0 or 1. This brings in some problems for assessing goodness-of-fit, as will be discussed below.

Assuming that $L(\underline{\Theta})$ is well behaved, it is possible to find a unique set of estimates of $\underline{\Theta}$, $\hat{\underline{\Theta}}$, which maximises $L(\hat{\underline{\Theta}})$ where $\hat{\underline{\Theta}}$ depends

on the observations. If we define

$$\ell(\underline{\theta}) = \ell n\, L(\underline{\theta}) \qquad (13)$$

and

$$\underline{\underline{V}} = -\{E\left(\frac{\partial^2 \ell(\underline{\theta})}{\partial \underline{\theta}^2}\right)\}^{-1} \qquad (14)$$

where $E(\cdot)$ denotes an expectation operator(*), then $\underline{\hat{\theta}}$ is an asymptotically efficient estimator of $\underline{\theta}$ and is asymptotically distributed as Normal, $N(\underline{\theta},\underline{\underline{V}})$. Moreover $-2.\ell(\underline{\hat{\theta}})$ is asymptotically distributed χ^2 (chi-squared) with Q degrees of freedom. This means that although $\underline{\hat{\theta}}$ may be biased for small samples, the bias is small for large enough Q (just how large is 'large enough' is a function of the problem under examination, but generally data sets with 500 to 1000 observations have been found to be sufficient). The estimator $\underline{\hat{\theta}}$ is the best possible for large samples (McFadden, 1976), and there is a concrete expression $\underline{\underline{V}}$ for its variance-covariance matrix. Note however, that for most model forms, including the easy to handle MNL, $\underline{\hat{\theta}}$ must be calculated by an iterative procedure. Fortunately $\underline{\underline{V}}$ is useful in this iterative calculation and is thus available when convergence occurs.

For a simple MNL model of the form

$$P_i(\underline{A}(q),\underline{\theta}) = \frac{\exp(\underline{Z}_{iq}.\underline{\theta})}{\sum_{j\epsilon\underline{A}(q)} \exp(\underline{Z}_{jq}.\underline{\theta})} \qquad (15)$$

... ...

(*) For the simple MNL model the expectation is not needed because the second derivatives of $\ell(\underline{\theta})$ depend only on the modelled probabilities and not on the observed proportions or choices (see, Hartley and Ortuzar, 1980).

the Likelihood Function can be written as follows (Ben-Akiva, 1973)

$$L(\underline{\theta}) = \prod_{q=1}^{Q} \prod_{j \in \underline{A}(q)} P_j(\underline{A}(q), \underline{\theta})^{g_{jq}} \qquad (16)$$

where g_{jq} equals 1 if alternative j was selected in observation q and zero otherwise. Taking the natural logarithm of both sides we get

$$\ell(\underline{\theta}) = \sum_{q=1}^{Q} \sum_{j \in \underline{A}(q)} g_{jq} . \ln P_i(\underline{A}(q), \underline{\theta}) \qquad (17)$$

Substituting equation (15) in (17) we can derive the first order conditions (McFadden, 1974)

$$\frac{\partial \ell(\underline{\theta})}{\partial \theta_k} = \sum_{q=1}^{Q} \sum_{j \in \underline{A}(q)} \{g_{jq} - P_i(\underline{A}(q), \underline{\theta})\} . z_{jq}^k = 0 \qquad (18)$$

for k=1, ..., K

It is easy to see that if the set of variables includes a mode-specific dummy as follows:

$$z_{jq}^k = \begin{cases} 1 & \text{for } j = a \\ & \\ 0 & \text{otherwise} \end{cases} \qquad (19)$$

then from the first order conditions (18) we will always get

$$\sum_{q=1}^{Q} g_a = \sum_{q=1}^{Q} P_a(\underline{A}(q), \underline{\theta}) \qquad (20)$$

Therefore a comparison of a sum of probabilities for a given alternative with the total number of observations that selected the alternative can be misleading. For this reason, and because it is also misleading to compare the computed probabilities with

the g_{jq} variables (if we assume that the actual choice is made with a probability and not a certainty as the g_{jq} variables indicate), a goodness-of-fit measure such as R^2 in ordinary least squares, which is based on estimated residuals, does not exist.

A word of caution is also in order here. Although it is well known that for a logit model with linear-in-parameters specification $\ell(\underline{\theta})$ is well behaved, this has not been proven for probit models, except for the simplest independent binary case. Indeed it has been noted that the most widely used and efficient MNP estimation computer code available, CHOMP (Daganzo and Shoenfeld, 1978) may have problems in that the approximation to $\ell(\underline{\theta})$ used is not necessarily unimodal (Bouthelier, 1978; Daganzo, 1979).

The well understood properties of the maximum likelihood estimation method, for well behaved likelihood functions, allow a number of statistical tests which are of major importance:

(i) *The t-test for significance of any component* $\hat{\theta}_k$ *of* $\underline{\hat{\theta}}$

Equation (14) implies that $\hat{\theta}_k$ has an estimated variance v_{kk}, where $\underline{\underline{V}} = \{v_{kk}\}$, which is calculated by the estimating program. Thus if $\theta_k = 0$,

$$t = \hat{\theta}_k / v_{kk} \qquad (21)$$

is distributed Normal $N(0,1)$. For this reason, it is possible to test whether $\hat{\theta}_k$ is significantly different from zero (it is not exactly a t-test as this is a large sample approximation; t is tested with the Normal distribution). Large absolute values of t (typically bigger than 2 for 95% confidence levels) lead to the rejection of the null hypothesis $\theta_k = 0$ and hence to acceptance that $\hat{\theta}_k$ is significant.

(ii) *The likelihood ratio test of linear restrictions of any general hypothesis*

A number of important model properties can be expressed as linear restrictions on a more general linear-in-parameters model. Some important examples of properties are:

- Attribute genericity: There are two main types of explanatory variables, 'generic variables' and 'alternative-specific' variables. The former vary in value (or level) across choice alternatives, whereas the latter are those with an identifiable correspondence between choice alternatives; because they may not vary across all options, alternative-specific-variables can take on a zero value for certain elements of the choice set. Let us assume a model with three alternatives, car, bus and rail, and the following choice influencing variables.

TT = travel time OPC = out-of-pocket travel costs

Then, a general form of the model would be:

$$\overline{U}_{car} = \theta_1 OPC_{car} + \theta_2 TT_{car}$$

$$\overline{U}_{bus} = \theta_3 OPC_{bus} + \theta_4 TT_{bus}$$

$$\overline{U}_{rail} = \theta_5 OPC_{rail} + \theta_6 TT_{rail}$$

However, it might be hypothesised that costs (but not times, say) should be generic. This can be expressed by writing this hypothesis as two linear equations in the parameters:

$$\theta_3 - \theta_1 = 0$$

$$\theta_5 - \theta_1 = 0$$

In general it is possible to express attribute genericity by linear restrictions on a more general model. For extensive use of this type of test refer to Dehghani and Talvitie (1979).

- Sample homogeneity: It is possible to test whether or not the same model coefficients are appropriate for two subpopulations (say living north and south of a river). For this, one formulates a general model using different coefficients for the two populations and then tests for equality of the coefficients as a linear restriction.

Because of the properties of ML, it is very easy to test any such hypothesis expressed as linear restrictions by means of the well-known *likelihood ratio* test (LR). To perform the test the estimation program is first run in the more general case to give the estimates $\hat{\underline{\theta}}$ and the log-likelihood at convergence $\ell^*(\hat{\underline{\theta}})$. It is then run again to attain estimates $\hat{\underline{\theta}}_r$ of $\underline{\theta}$ and the new log-likelihood at maximum $\ell^*(\hat{\underline{\theta}}_r)$, for the restricted case. Then, if the restricted model under consideration is a correct specification the likelihood ratio statistic,

$$-2\{\ell^*(\hat{\underline{\theta}}_r) - \ell^*(\hat{\underline{\theta}})\}$$

is asymptotically distributed χ^2 with k-r degrees of freedom where k is the number of elements in $\underline{\theta}$ and r is the number of linear restrictions(*). Rejection of the null hypothesis implies that the restricted model is erroneous. Train (1977), offers examples of the use of this test to study questions of non-linearity, non-genericity and non-homogeneity. Horowitz (1980) has discussed the power and properties of the test in detail and should be consulted for further reference.

... ...

(*) Note that for this we need one model to be a restricted or nested version of the other. We will look at what to do with non-nested models below.

(iii) *The overall test of fit*

A special case of the LR test is to find out whether all components of $\underline{\hat{\theta}}$ are equal to zero - the equally likely model:

$$P_i(\underline{A}(q),\underline{\theta}) = \frac{1}{N_q} \qquad (22)$$

where N_q is the number of options available to individual q ; or, preferably, to test whether those components of $\underline{\hat{\theta}}$ which do not correspond to model constants are equal to zero - the 'best null' model (or 'constants only' model):

$$P_i(\underline{A}(q),\underline{\theta}) = ms_i \qquad (23)$$

where ms_i is the market share of alternative i. Let us consider the first case, which is the most common and obvious one, to begin with.

If there are K parameters and $\ell^*(0)$ is the log-likelihood of the equally-likely model, then under the null hypothesis of $\underline{\theta} = 0$, the value

$$-2\{\ell^*(0) - \ell^*(\underline{\hat{\theta}})\}$$

should be asymptotically distributed χ^2 with k degrees of freedom. Note that $\ell^*(0)$ does not require a special program run since it is usually calculated as the initial log-likelihood at the start of the program. This test is actually rather weak; if rejected it only says that the model with parameters $\underline{\hat{\theta}}$ provides a better explanation of the data than a model which does not have any significant explanatory power (the equally likely model). It is obvious that when the model contains alternative-specific constants, the test in this simplest form is not appropriate. It is more relevant to test, as suggested above, whether the explanatory

variables add anything to the explanation given by the constants alone, ie. the best null model. It is rather embarrasing to note that constants tend to account for 60% to 80% of the explanatory power of these models (Talvitie and Kirschner, 1978).

In general, an extra run is required to calculate $\ell^*(C)$, the log-likelihood of the model containing only alternative-specific constants, except for models when all individuals face the same alternatives where it has the following close form equation (Tardiff, 1976a).

$$\ell^*(C) = \sum_{j=1}^{J} Q_j \ln \frac{Q_j}{Q} \qquad (24)$$

where Q_j = number of individuals choosing alternative A_j.

(iv) *The Rho squared indices*

It is felt by many that a coefficient of goodness-of-fit is useful. However, as we mentioned above, a goodness-of-fit like R^2 in ordinary least squares does not exist. A goodness-of-fit coefficient should range from 0 to 1 (no fit to perfect fit), be meaningful for comparing models calibrated with different samples, and hopefully be related to a statistic with a known probability distribution for purposes of statistical hypothesis testing.

Such an index has been defined (McFadden, 1976) as

$$\rho^2 = 1 - \frac{\ell^*(\hat{\underline{\theta}})}{\ell^*(0)} \qquad (25)$$

However, it has been noted that although ρ^2 behaves nicely at the limits (eg. 0 and 1) it does not have an intuitive interpretation between the limits (Hauser, 1978). A quotation by McFadden (1976) may also be appropriate at this point:

> "... Those unfamiliar with the ρ^2 should be forewarned that its values tend to be considerably lower than those of the R^2 index (of regression analysis) and should not be judged by the standards for "good fit" in ordinary regression analysis. For example, values of 0.2 to 0.4 for ρ^2 represent an excellent fit ..."

Because a ρ^2-like index can in principle be computed relative to any null hypothesis, it is important to choose an appropriate one. For example, it is very easy to show that the minimum values of ρ^2 (with respect to the equally likely model), in models with alternative-specific constants, vary depending on the proportion of individuals choosing each alternative. Taking a simple binary case, Table 1 (Tardiff, 1976) show the minimum values of ρ^2 for different proportions choosing option 1. It can be seen that ρ^2 is only appropriate for the 50/50 percent case.

Sample Proportion Selecting the First Alternative	Minimum value of ρ^2
0.50	0.00
0.60	0.03
0.70	0.12
0.80	0.28
0.90	0.53
0.95	0.71

Table 1. Minimum values of ρ^2 for various relative frequencies

(Source: Tardiff, 1976)

These values mean, for example, that a model calibrated with a 0.9/0.1 sample, yielding a ρ^2 of 0.55 would undoubtedly be much weaker than a model yielding a ρ^2 of 0.25 from a sample with a 0.5/0.5 split. Fortunately, a rather simple adjustment exists

(Tardiff, 1976) that overcomes these difficulties. It consists of defining a more appropriate index $\bar{\rho}^2$ as

$$\bar{\rho}^2 = 1 - \frac{\ell^*(\hat{\underline{\theta}})}{\ell^*(C)} \qquad (26)$$

This statistic has between 0 and 1, is comparable across different samples and is also related to the χ^2 statistic; therefore it is recommended over ρ^2. (For a more profound discussion of these issues, the reader is referred to the recent papers by Gunn and Bates, 1980; and Horowitz, 1980)

5.3 Model comparison through goodness-of-fit measures

It has been shown (see, for example Horowitz, 1980) that uncritical use of goodness-of-fit statistics, such as ρ^2, can give perverse results(*). For this reason, among others, several other possible measures have been proposed and discussed by, for example, Stopher (1975); McFadden (1976); and Hauser (1978). We will, however, mention only one other measure, the 'first preference recovery', FPR (also termed the 'percentage correctly predicted' or 'percent right' for short) and discuss a recent improvement to it (Gunn and Bates, 1980). FPR is an aggregate measure which simply computes the proportion of individuals that actually select the option with the highest modelled utility. FPR is easy to understand and can readily be compared to the 'chance recovery', CR, the recoveries predicted using the equally likely model, given by:

... ...

(*) Especially if one is comparing non-nested models.

$$CR = \frac{\sum_{q=1}^{Q} 1/N_q}{Q} \qquad (27a)$$

or, if every individual has the same number of options $\bar{N}$, by:

$$CR = 1/\bar{N} \qquad (27b)$$

FPR can also be compared to the 'market share recovery', MSR, the recoveries predicted by the best null model (constants only model given by:

$$MSR = \sum_j (ms_j^2) \qquad (28)$$

where ms_j = market share of option j.

Also, being an aggregate test it has strong intuitive appeal and is useful to improve communication between analysts and managers or decision-makers (Hauser, 1978). Unfortunately, because of its aggregate nature, it can be misleading. For example

> "... a first preference recovery of 55% is usually good, but not in a market of two products. A recovery of 90% is usually good in a two-product market but not if one product has a market share of 95%." (Hauser, 1978)

Two further problems of FPR, in the sense of not being an unambiguous indicator of model reliability, are worth noting. The first is that too high a value of FPR should lead to rejecting the model as well as too low. To understand this point is is necessary to define the expected value of FPR for a specified model. This is given by

$$ER = \sum_q p_q \qquad (29)$$

where p_q is the calculated (maximum) probability associated with the best option for individual q. We also need to note that the variance of CR and ER are given respectively by(*)

$$\text{Var(CR)} = \sum_{q=1}^{Q} \frac{1}{N_q}\left(1 - \frac{1}{N_q}\right) \qquad (30)$$

and

$$\text{Var(ER)} = \sum_{q=1}^{Q} p_q(1 - p_q) \qquad (31)$$

Thus, a computed value of FPR for a given model can be compared with CR and ER; if the three measures are relatively close (given the estimated variances) the model is *reasonable but uninformative*; if FPR and ER are similar and larger than CR, the model is *reasonable and informative*; if FPR and ER are not similar, the model *does not* explain the variation in the data and should be rejected - *whether FPR is larger or smaller than ER.*

The second problem with the measure arises even if the value of FPR is acceptable, because a test which weights each correct prediction equally will not be suitable for circumstances where some options are more important than others. For example, given a multimodal choice context if we are particularly interested in the predictions with respect to a minor mode, say, park-and-ride (P&R), we would not judge two models with the same FPR equivalent, if one of them predicted P&R incorrectly in all cases whilst correctly predicting the other modes slightly more times than the rival model which performed reasonably well for all modes.

... ...

(*) Because for an individual q, and FPR is an independent random event occurring with probability $1/N_q$ and p_q respectively.

Gunn(*) has obtained a more sensitive test based on the abovementioned measure by extending the comparison of observed and expected FPR to take account of 'where they occur' as well as their absolute number. For this he divides the probability range (0,1) into a number of intervals - for example (0,0.1), (0.1,0.2), ..., (0.9,1.0) - and allocates individual observations to each of these intervals on the basis of their modelled 'first-preference probabilities' (fpp) (ie. the highest probability predicted by the model). Thus, if two individuals have, respectively say, fpp = 0.488 and 0.415, they would both be assigned to the interval (0.4,0.5). On the basis of the model, we can expect approximately 45% of these individuals to show FPR. We can, then, observe the actual number of FPR in that group and compare expectation (on the basis of the model) with out-turn. It is interesting to realise that a given model might have exactly the expected number of FPR overall and yet be incorrect in the distribution of FPR over the spectrum between likely and unlikely recovery. It is obvious that this would indicate a faulty model structure as clearly as an incorrect overall number of FPR. Comparisons between observed and expected frequencies can be carried out by means of straightforward χ^2 tests (see Ortuzar, 1980c).

5.4 Validation samples

The performance of any model should be judged against data other than that being used to specify it and, ideally, taken at another point in time (perhaps after the introduction of a policy in order to judge the model response properties). This is most

... ...

(*) Private communication to be written as a Technical Note, Institute for Transport Studies. Examples of its use are given in Gunn and Bates (1980); and Ortuzar (1980c).

obviously true for the sort of models (eg. gravity model) frequently fitted to aggregate data sets, because a comparison of such models to the calibration data can only reveal how good a summary they provide for that one data set. The same is true though of disaggregate models. We will define a subsample of the data, or preferably another sample, *not used during estimation*, as a validation sample.

In this section we will describe a procedure to estimate the minimum size of such a validation sample (to be subtracted from the total sample available for the study) conditional on allowing us to detect a difference between the performance of two or more models, when there exists a true difference between them. The method, due to Gunn, is based on the FPR concept and will be used elsewhere to determine the size of a validation sample for the estimation of disaggregate choice models (Ortuzar, 1980c).

Consider a 2x2 table layout as follows:

		Model 2	
		Not FPR	FPR
Model 1	Not FPR	n_{11}	n_{12}
	FPR	n_{21}	n_{22}

n_{ij} = number of individuals assigned to cell (i,j)

For all individuals in a validation sample, choice probabilities and FPR are calculated for each of two models under investigation and the cells of the table are filled appropriately, for example assigning to cell (1,1) if not FPR in either model, etc.

We are interested in the null hypothesis that the probabilities with which individuals fall into cells (1,2) and (2,1) are equal, for in that case the implication, on simple FPR, is that the two models are equivalent. On this null hypothesis, the statistic M (after McNemar, see Foerster, 1979a)

$$M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \qquad (32)$$

is χ^2 distributed with 1 degree of freedom. Thus, a test of the 'equivalence' of the two models, in terms of FPR, is given by computing M and comparing the result with $\chi^2_{\alpha,1}$. If M is less than the appropriate chosen critical value of $\chi^2_{\alpha,1}$ (3.85 for the usual 95% confidence level) we cannot reject the null hypothesis and we conclude the models are equivalent in these terms.

Given this procedure we can choose whichever level of confidence seems appropriate for the assertion that the two models under comparison differ in respect of the expected number of FPR. This gives us control over the fraction of times that we will incorrectly assert a difference between similar models. As usual, the aim of selecting a particular sample size is to ensure a corresponding control over the proportion of times we will make the other type of error, namely incorrectly concluding that there is no difference between different models. Now, to calculate the probability of an error of the second type we need to decide what should be the minimum difference that we should like to be able to detect. With this we can calculate the sample size needed to reduce the chance of errors of the second kind to an acceptable level for models which differ by exactly this minimum amount, or more.

Consider, as an illustration, a particular case of two models such that, on average, model 2 produces 10 *extra* FPR per 100 individuals modelled as compared to model 1(*). In this simple case n_{21} is zero and the statistics M simply becomes n_{12}. If we are ensuring 95% confidence that any difference we establish could not have arisen by chance from equivalent models, we will compare n_{12} with the value 3.85. For any given sample size n, say, the probability that r individuals will be assigned to cell (1,2) is simply the binomial probability $\binom{n}{r}p^r(1-p)^{n-r}$ where p denotes the probability of an individual chosen at random being assigned to the (1,2) cell, eg. the minimum difference we have set to detect. Given n, and taking p = 0.05, say, we can calculate the probabilities of 0,1,2 and 3 individuals being assigned, and sum these to give the total probability of accepting the null hypothesis, eg. committing an error of the second kind. Table 2 gives the resulting probabilities for different sample sizes(†). It is clear that the required validation sample size needs to be relatively large, given that estimation data sets are only a few hundred data points. Also recall that this table is for the simple case of one model being better than or equal in each

... ...

(*) Note that here it does not matter whether this arises as a result of model 1 having 20% FPR and model 2, 30% FPR, or model 1 80% FPR and model 2, 90%; in other words *both* models can be inadequate.

(†) An extension of this table for other values of p is given in Chapter 7 of Ortuzar (1980c).

Sample size	Minimum difference 5%
	Prob {error II}
50	0.76
100	0.26
150	0.05
200	0.01
250	0.00

Table 2. Probability of an error of the second kind for given sample size, minimum difference of 5%, and models as defined

observation than the other, although the method is easily generalisable to cases where both (1,2) and (2,1) cells have non-zero probability.

5.5 Comparison of non-nested models

The likelihood-ratio tests outlined in section 5.2 above, require testing a model against a parametric generalisation of itself, ie., it requires the models to be 'nested'. Models whose utility functions have significantly different functional forms or models based on different behavioural paradigms cannot be compared by these tests.

It is easy to conceive of situations in which it would be useful to test a given model against another which is not a parametric generalisation of itself. The following example provided by Horowitz (1980) is very illustrative. Suppose that one model has a representative utility function specified as:

$$\overline{U} = \theta_1 z_1 + \theta_2 z_2 \qquad (33)$$

and the other, a representative utility function given by:

$$\overline{V} = \theta_3 z_3 z_4 \qquad (34)$$

and that it is desired to test the two models against one another to determine which best explains the data. Clearly there is no value of θ_3 that causes $\overline{U}$ and $\overline{V}$ to coincide for all values of θ_1, θ_2 and the attributes $\underline{z}$.

If both models belong to the same general family of models, it is possible to construct hybrid models; for instance, in our simple example we could form a model whose representative utility $\overline{W}$ contains both $\overline{U}$ and $\overline{V}$ as special cases:

$$\overline{W} = \theta_1 z_1 + \theta_2 z_2 + \theta_3 z_3 z_4 \qquad (35)$$

Using likelihood-ratio tests, both models can be compared against the hybrid. The first (33) corresponds to the hypothesis $\theta_3 = 0$ and the second (34) to the hypothesis $\theta_1 = \theta_2 = 0$. Several other tests, including cases where the competing models do not belong to the same general family are discussed at length in the excellent paper by Horowitz (1980).

An especially helpful feature of the validation sample concept discussed in section 5.4 above, is that, provided the sample is adequate, the issue of ranking models, nested or non-nested, is particularly easily resolved (Gunn and Bates, 1980), because likelihood ratio tests can be performed on that sample for any models regardless of difference in model structure or parameters(*).

... ...

(*) The condition of one model being a parametric generalisation of the other is only required for tests with the *same* data used for estimation (Gunn and Bates, 1980).

5.6 Estimation of models from choice-based samples

We mentioned in section 2, that estimating a choice model from a choice-based sample may be of interest because the data collection costs are often considerably smaller than those for typical random or stratified samples (Lerman and Manski, 1976; 1979). The problem of finding a tractable estimation procedure possessing certain desirable statistical properties, is not an easy one; the state-of-the-art is provided by the excellent papers of Manski and Lerman (1977) and Manski and McFadden (1980).

These authors have found that appropriate maximum likelihood estimators for choice based sampling, *except in very restricted circumstances* are impractical due to computational intractability. However, if it is assumed that the analyst knows the fraction of the decision-making population selecting each alternative then a tractable method can be introduced. This approach modifies the familiar maximum likelihood estimator of random sampling by weighting each observation's contribution to the log-likelihood by the ratio H(i)/S(i), where H(i) is the fraction of the population selecting option i and S(i) is the analogous fraction for the choice-based sample. Manski and Lerman (1977) go on and prove that this estimator is consistent, find its asymptotic covariance matrix and examine its asymptotic efficiency for special cases. They also show that the unweighted random sample ML estimator is generally inconsistent when applied to choice-based samples, and in most choice models this inconsistency affects all parameter estimates. However, for simple MNL models with a *full set* of alternative-specific dummy variables, *the inconsistency is fully confined* to the estimates of the coefficients of these dummies (Manski and Lerman, 1977; Manski and McFadden, 1980).

This latter result has been used in an empirical study in South Africa by Stopher and Wilmot (1979). Coslett (1980) have extended this work to the estimation of hierarchical logit models discussed below.

5.7 Estimation of hierarchical logit models

The nested or hierarchical logit model (Williams 1977; Daly and Zachary, 1978) is a generalisation of the MNL which does not suffer the 'independence from irrelevant alternatives' restriction. For example, if we consider the well-known red bus/blue bus case, a hierarchical logit model would proceed in two stages. Firstly, a primary split between car (c) and 'composite' bus mode (b), and secondly a subsplit between the two bus options (rb and bb, respectively), as shown in Figure 3. A detailed description of the calibration and properties of such a model, for choice among car, bus and train, using aggregate data has been presented in Hartley and Ortuzar (1980). Here we just want to show the special complications that arise when the estimation is carried out using individual choice data. For practical examples refer to Coslett (1980), Sobel (1980), and Ortuzar (1980c).

Individuals are conceptually assumed to evaluate each alternative according to utility functions U_c, U_{rb} and U_{bb} respectively (with measurable components $\bar{U}_c$, $\bar{U}_{rb}$ and $\bar{U}_{bb}$) as in the case of the MNL. However, in this case we need also to consider a 'composite utility' of the lower hierarchy or 'nest'. This composite utility ($\bar{U}_b$) includes the expected value of the maximum utility of the members of the next, given by

$$I_b = \ln \{\exp(\bar{U}_{rb}) + \exp(\bar{U}_{bb})\} \qquad (36)$$

and attributes which are common to all the members of the lower hierarchy as in

$$\bar{U}_b = \phi I_b + \underline{\theta}\, \underline{z}_b \qquad (37)$$

where α is an estimated coefficient and $\underline{\theta}$ is the vector of estimated coefficients multiplying the set of attributes $\underline{z}_b$ which are common to all nest members(*).

... ...

(*) The reason for taking the attributes $\underline{z}_b$ out is that, being common, they do not influence the choice in the lower hierarchy (e.g. both buses have the same fare structure). However they must be included again in the next hierarchy because they certainly influence choice between car and the composite bus mode.

It is easy to see that the hierarchical logit model can be estimated using standard MNL software in two stages: firstly, as a binary logit model between red bus and blue bus, the results of which allow us to calculate I_b from (36); secondly this value is entered as another independent variable along with the $\underline{z}_b$ variables and the attributes of car in the primary split which is, in this simple case, another binary logit model. The secondary split will yield P(rb/b) and P(bb/b), the conditional probabilities of red bus and blue bus *given* that choice is constrained to bus. The primary split yields P(c) and P(b), the marginal probabilities of car and bus respectively. It is clear that probabilities of each mode are

$$\begin{aligned} P_{car} &= P(c) \\ P_{red\ bus} &= P(b).P(rb/b) \\ P_{blue\ bus} &= P(b).\ P(bb/b) \end{aligned} \qquad (38)$$

An important feature of the model concerns acceptable values of ϕ, the coefficient of the expected maximum utility of the nest (see Ortuzar, 1980b for a discussion of its use as a diagnostic tool for appropriate specification). Williams (1977) has shown that ϕ must satisfy:

$$0 < \phi \leq 1 \qquad (39)$$

it has also been shown (Williams, 1977; Daly and Zachary, 1978) that if there are more than two levels of nesting, eg. a case with more composite utilities and coefficients ϕ, then

$$0 < \phi_1 \leq \phi_2 \leq \phi_3 \leq \ldots \leq 1 \qquad (40)$$

where ϕ_1 represents the coefficient of the expected maximum utility of the 'lowest' hierarchy. Note also, that at any hierarchical level, i, a value of $\phi_i = 1$ implies that the linked nesting at level i is mathematically equivalent to a simple MNL at that level. For a good discussion of these issues see Coslett (1980) and for a review and an application to real data see Ortuzar (1980) and Sobel (1980), who has shown that for hierarchical logit models there exist equivalent measures to the ρ^2 and $\bar{\rho}^2$ indices (equations 25 and 26), given by

$$\rho^2 = 1 - \frac{\ell_1^*(\hat{\underline{\theta}}) + \ell_2^*(\hat{\underline{\theta}}) + \ldots + \ell_j^*(\hat{\underline{\theta}})}{\ell_1^*(0) + \ell_2^*(0) + \ldots + \ell_j^*(0)} \qquad (41)$$

and

$$\bar{\rho}^2 = 1 - \frac{\ell_1^*(\hat{\underline{\theta}}) + \ell_2^*(\hat{\underline{\theta}}) + \ldots + \ell_j^*(\hat{\underline{\theta}})}{\ell_1^*(C) + \ell_2^*(C) + \ldots + \ell_j^*(C)} \qquad (42)$$

where the subscripts 1 to j refer to the simple MNL models in the hierarchy of interest.

Notwithstanding the simplicity of the 'heuristic' or 'bottom up' calibration of the hierarchical logit model (Williams, 1977) it is known that the consequence of sequential estimation is a loss of statistical efficiency which may be severe (Daly and Zachary, 1978; Amemiya, 1976, 1978; Coslett, 1980; Sobel, 1980). This happens because the standard errors of lower level coefficient estimates permeate from lower hierarchies upwards imbedded in the values of the expected maximum utilities I. When there are multiple hierarchies,

> "... successively higher level expected maximum utilities will contain greater and greater proportions of random statistical 'noise'."
> (Sobel, 1980)

What is really required is a simultaneous estimation routine which would eliminate the compounding effect of these errors, thereby improving the statistical efficiency of the estimates of the parameters ϕ(*). Another powerful reason for developing such software is to avoid the unpleasant possibility of obtaining different estimates of the same parameter at different hierarchical levels (which is quite common due to the different amount and quality of data used in each). At least two experimental simultaneous estimation software packages are in the process of development by Daly at Cambridge Systematics Inc. and by Small and Brownstone of Princeton University, but none is yet available.

ACKNOWLEDGEMENTS

I am grateful to Hugh Gunn, Dirck Van Vliet and Huw Williams for all they have taught me, part of which is reflected heavily in this paper.

... ...

(*) Recall how crucial are the ϕ's in allowing for structural diagnosis of the model, through conditions (39) and (40).

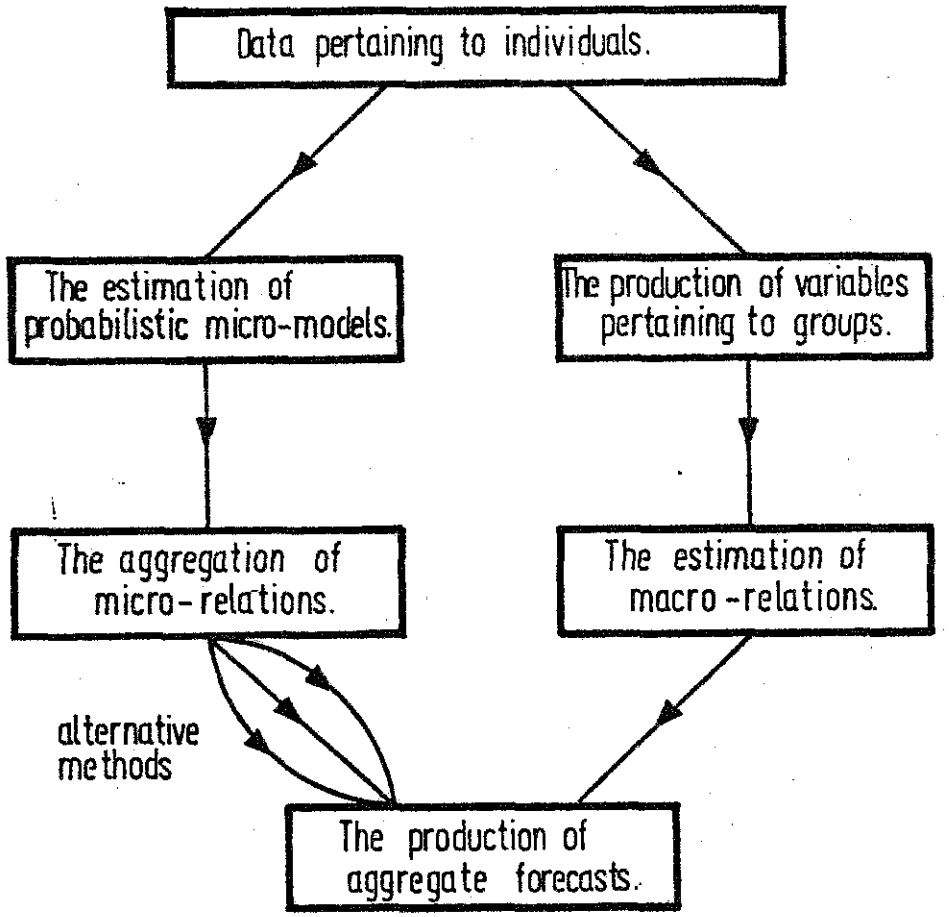

FIGURE 1a: Alternative aggregation strategies.

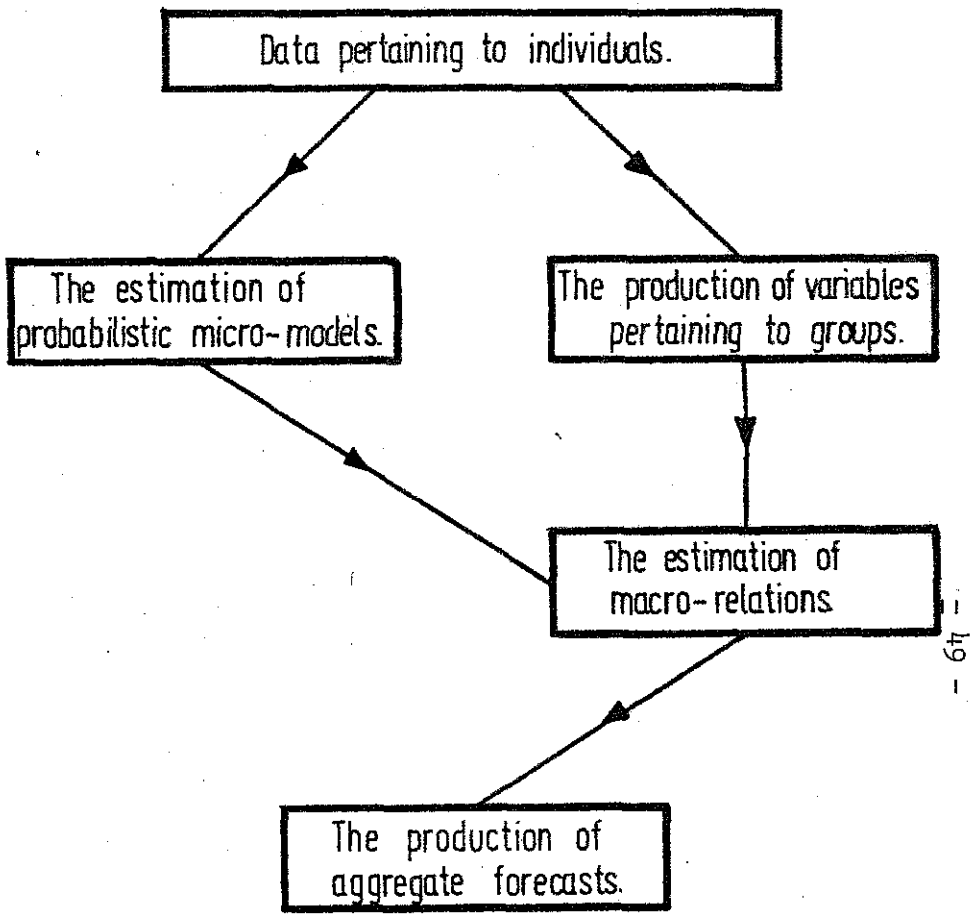

FIGURE 1b: The aggregation strategy implicit in British Studies.

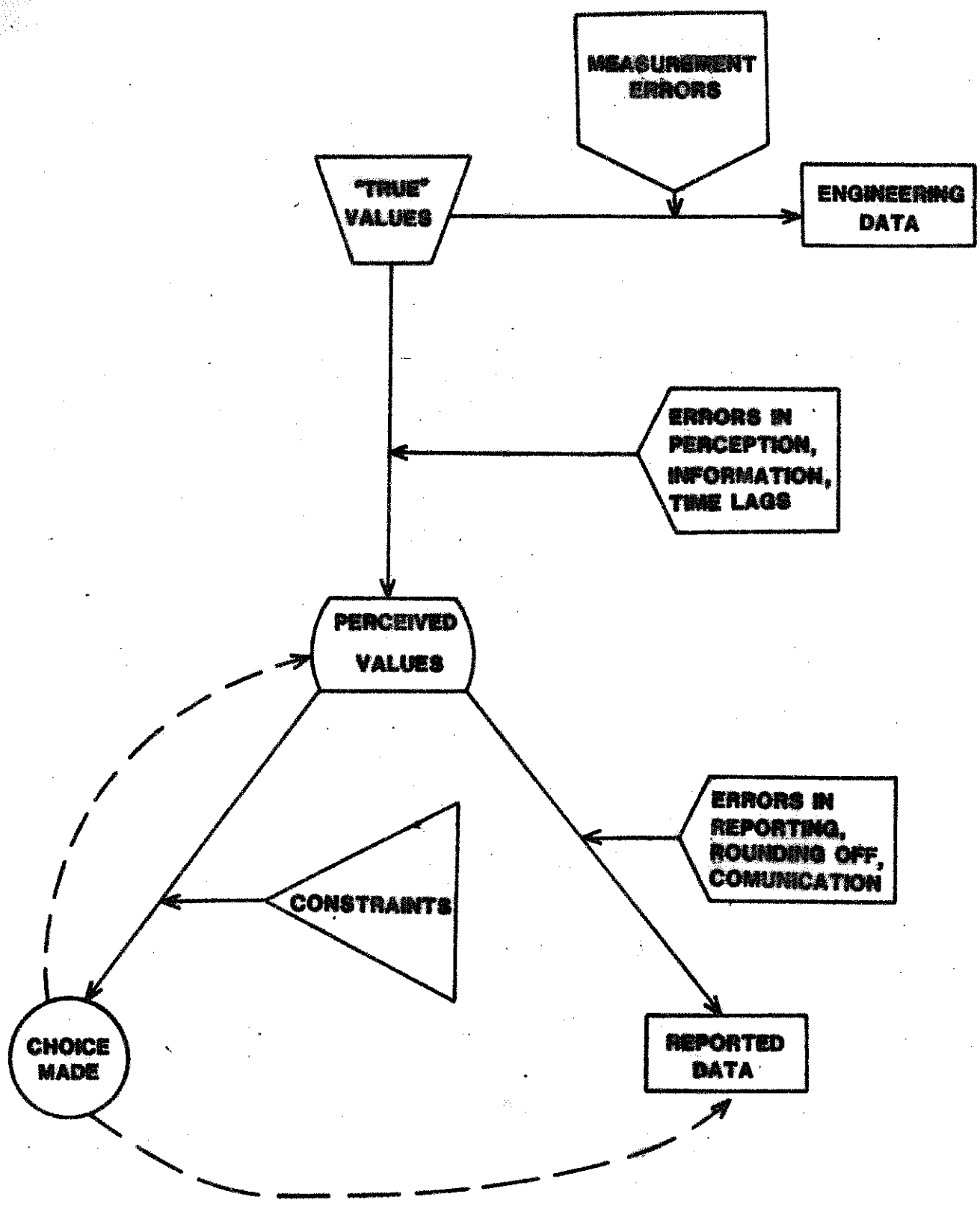

FIGURE 2: NOTIONAL RELATIONSHIP BETWEEN CHOICE AND DIFFERENT ATTRIBUTE MEASUREMENTS

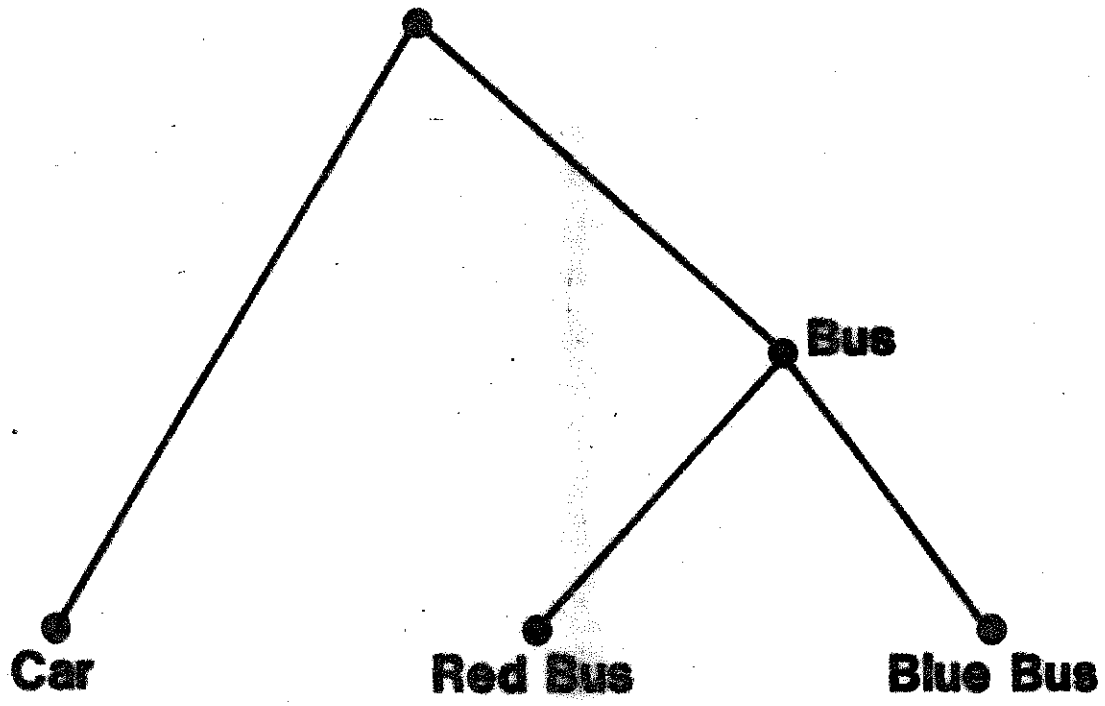

FIGURE 3: A SIMPLE HIERARCHICAL LOGIT MODEL

REFERENCES

AMEMIYA, T. (1976) "The specification and estimation of a multivariate logit model" Stanford: Stanford University, Institute for Mathematical Studies in the Social Sciences, Tech.Report 211.

AMEMIYA, T. (1978) "On a two-step estimation of a multivariate logit model", Journal of Econometrics, Vol.8, pp.13-21.

BEN-AKIVA, M.E. (1973) "Program for maximum likelihood estimation of the multinomial logit model", Cambridge: Massachusetts Institute of Technology, Department of Civil Engineering, Working Paper (unpublished).

BEN-AKIVA, M.E. (1979) "Issues in transferring and updating travel behaviour models", Fourth International Conference on Behavioural Travel Modelling 1979, Eibsee, July 1979.

BEN-AKIVA, M.E., ADLER, T.J., JACOBSON, J. and MANHEIM, M.L. (1977) "Experiments to clarify priorities in urban travel forecasting research and development", Final Report to the Office of University Research, U.S. Department of Transportation, by the Center for Transportation Studies, Massachusetts Institute of Technology, CTS Report Number 77-24.

BEN-AKIVA,M.E. and ATHERTON, T.J. (1977) "Methodology for short-range travel demand predictions: analysis of carpooling incentives", Journal of Transport Economics and Policy, Vol.11, No.3, pp. 224-261.

BEN-AKIVA, M.E. and KOPPELMAN, F.S. (1974) "Multidimensional choice models: alternative structures of travel demand models", Transportation Research Board, Special Report 149, pp.129-142.

BEN-AKIVA, M.E., LERMAN, S.R., DAMM, D., JACOBSON, J. and WEISBROD, G (1979) "Understanding, prediction and evaluation of transportation related consumer behaviour", Phase I Report to the Office of University Research, U.S. Department of Transportation, Washington D.C., by the Center for Transportation Studies, Massachusetts Institute of Technology.

BISHOP, Y.M., FIENBERG, S.E. and HOLLAND, P.W. (1975) Discrete Multivariate Analysis - Theory and Practice. Cambridge: M.I.T. Press.

BOUTHELIER, F.J. (1978) "An efficient methodology to estimate and predict with multinomial probit models: applications to transportation problems", Cambridge: Massachusetts Institute of Technology, Department of Ocean Engineering, PhD Thesis (unpublished).

BOUTHELIER, F.J. and DAGANZO, C.F. (1979) "Aggregation with multinomial probit and estimation of disaggregate models with aggregate data: a new methodological approach", Transportation Research, Vol. 13B, No.2, pp. 133-146.

BOYCE, D.E., DESFOR, G., et al. (1974) "Impact of suburban rapid transit station location, fare and parking availability on users' station choice behaviour: analysis of the Philadelphia-Lindelwold high-speed line", Final Report to the Office of the Secretary, U.S. Department of Transportation, Washington D.C., Contract DOT-05-10044, by the Regional Science Department, University of Pennsylvania.

BRUZELIUS, N. (1979) The value of Travel Time. Theory and Measurement. London: Croom Helm.

CARDELL, N.S. and REDDY, B.J. (1977) "A multinomial logit model which permits variations in tastes across individuals", Charles River Associates Inc. (unpublished).

COSSLETT, S. (1980) "Efficient estimation of discrete choice models", in C.F. Manski and D. McFadden (Eds.) Structural Analysis of Discrete Data: with Econometric Applications. Cambridge: M.I.T. Press.

DAGANZO, C.F. (1979) "Calibration and prediction with random utility models: some recent advances and unresolved questions", Fourth International Conference on Behavioural Travel Modelling 1979, Eibsee, July 1979.

DAGANZO, C.F. (1980) Multinomial Probit - The Theory and its Applications to Demand Forecasting, New York: Academic Press Inc.

DAGANZO, C.F. and SCHOENFELD, L. (1978) "CHOMP user's manual", Berkeley: University of California, Institute of Transportation Studies, Research Report UCB-ITS-RR-78-7.

DAGENAIS, M.G., GAUDRY, M.J.I. and LIEM, T.C. (1980) "Multiple regression analysis with Box-Cox transformations and nonspherical residual errors: a transportation application", Montreal: Universite de Montreal, Centre de Recherche sur les Transport, Publication No. 166.

DALY, A. (1976) "Modal split models and aggregation", Reading: Local Government Operational Research Unit, LGORU Transportation Working Note 6.

DALY, A.J. (1978) "Issues in the estimation of journey attribute values", in D.A. Hensher and M.Q. Dalvi (Eds.) Determinants of Travel Choice. Sussex: Saxon House.

DALY, A.J. (1979) "Some issues in the implementation of advanced travel demand models", Fourth International Conference on Behavioural Travel Modelling, 1979, Eibsee, July 1979.

DALY, A.J. and ZACHARY, S. (1978) "Improved multiple choice models", in D.A. Hensher and M.Q. Dalvi (Eds.) Determinants of Travel Choice. Sussex: Saxon House.

DEHGHANI, Y. and TALVITIE, A. (1979) "Model specification, model aggregation and market segmentation in mode choice models. Some empirical evidence", Buffalo: State University of New York, Civil Engineering and Applied Sciences, Working Paper 782-06.

DOMENCICH, T.A. and McFADDEN, D. (1975) Urban Travel Demand: A Behavioural Analysis. Amsterdam: North Holland.

FOERSTER, J.F. (1979a) "Mode choice decision process models: a comparison of compensatory and non-compensatory structures", Transportation Research, Vol. 13A, No. 1, pp. 17-28.

FOERSTER, J.F. (1979b) "Non-linear perceptual and choice functions: evidence and implications for analysis of travel behaviour", Fourth International Conference on Behavioural Travel Modelling 1979, Eibsee, July 1979.

GAUDRY, M.J.I. and WILLS, M.J. (1977) "Estimating the functional form of travel demand models", Montreal: Universite de Montreal, Centre de Recherche sur les Transport, Publication 63.

GENSCH, D.H. (1980) "Choice model calibrated on current behaviour predicts public response to new policies", Transportation Research, Vol. 14A, No. 2, pp 137-142.

GOODWIN, P.B. (1978) "On Grey's critique of generalised cost", Transportation, Vol. 7, No. 3, pp 281-295.

GUNN, H.F. and BATES, J.J. (1980) "Some statistical considerations in fitting travel demand models", International Conference on Research and Applications of Disaggregate Travel Demand Models, University of Leeds, 14-16 July, 1980.

GUNN, H.F., MACKIE, P.J. and ORTUZAR, J.D. (1980) "Assessing the value of travel time savings - a feasibility study on Humberside", Leeds: University of Leeds, Institute for Transport Studies, Working Paper (forthcoming).

HARTLEY, T.M. and ORTUZAR, J.D. (1980) "Aggregate modal split models: is current U.K. practice warranted?", Traffic Engineering and Control, Vol. 21, No. 1, pp 7-13.

HASAN, I. (1977) "A note on aggregate models", Berkeley: University of California, Department of Economics, Working Paper SL-7702.

HAUSER, J.R. (1978) "Testing the accuracy, usefulness, and significance of probabilistic choice models: an information-theoretic approach", Operations Research, Vol. 26, No.3, pp 406-421.

HAUSMAN, J.A. and WISE, D.A. (1978) "A conditional probit model for qualitative choice: discrete decisions recognising interdependence and heterogeneous preferences", Econometrica, Vol. 46, No. 2, pp 403-426.

HENSHER, D.A. (1972) "A study of modal choice and the value of travel time savings", Sidney: The University of New South Wales, Faculty of Commerce, School of Economics, PhD thesis (unpublished).

HENSHER, D.A. (1979a) "Behavioural demand modelling: some recent issues and possible new directions, in particular, choice set generation, endogenous attributes, decomposition of utility functions and congestion modelling", Regional Transport Research Workshop on Methods and Concepts in Transportation Modelling, Melbourne, January 1979.

HENSHER, D.A. (1979b) "Five contentions related to conceptual context in behavioural travel modelling", Fourth International Conference on Behavioural Travel Modelling 1979, Eibsee, July 1979.

HENSHER, D.A. (1979c) "Individual choice modelling with discrete commodities: theory and application to the Tasman bridge reopening", The Economic Records, September 1979, pp 243-260.

HENSHER, D.A. and JOHNSON, L.W. (1980) "Behavioural response and functional form", North Ryde: MacQuarie University, School of Economics and Financial Studies.

HENSHER, D.A. and LOUVIERE, J.J. (1979) "Behavioural intentions as predictors of very specific behaviour", Transportation, Vol.8, No. 2, pp. 167-182.

HOROWITZ, J. (1978) "The accuracy of the multinomial logit model as an approximation to the multinomial probit model of travel demand", Washington D.C.: U.S. Environmental Protection Agency (unpublished).

HOROWITZ, J. (1979a) "Identification and diagnosis of specification errors in the multinomial logit model", Washington D.C.: U.S. Environmental Protection Agency (unpublished)

HOROWITZ, J. (1979b) "Sources of error and uncertainty in behavioural travel demand models", Fourth International Conference on Behavioural Travel Modelling 1979, Eibsee, July 1979.

HOROWITZ, J. (1980) "Specification tests for probabilistic choice models", International Conference on Research and Applications of Disaggregate Travel Demand Models, University of Leeds, 14-16 July, 1980.

JANSEN, G.R.M., BOVY, P.H.L., VAN EST, J.P.J.M. and LE CLERCQ, F. (Eds.) (1979) New Developments in Modelling Travel Demand and Urban Systems. Westmead, Farnborough: Saxon House.

JOHNSON, L.W. and HENSHER, D.A. (1980) "Application of multinomial probit to a two period panel data set". International Conference on Research and Applications of Disaggregate Travel Demand Models, University of Leeds, 14-16 July, 1980.

JOHNSON, M. (1975) "Going to work by car, bus or BART: Attitudes, perceptions and decisions", Berkeley: University of California Institute of Transportation Studies, Working Paper 7518.

KOPPELMAN, F.S. (1974) "Prediction with disaggregate models: the aggregation issue", Transportation Research Record 527, pp 73-80.

KOPPELMAN, F.S. (1976a) "Guidelines for aggregate travel prediction using disaggregate choice models", Transportation Research Record 610, pp 19-24.

KOPPELMAN, F.S. (1976b) "Methodology for analysing errors in prediction with disaggregate choice models", Transportation Research Record 592, pp 17-23.

LEAMER, E. (1978) Specification Searches: Ad Hoc Inference with Non-experimental Data. New York: Wiley.

LERMAN, S.R. and LOUVIERE, J. (1978) "The use of functional measurement to identify the form of utility functions in travel demand models", Transportation Research Record 673, pp 78-85.

LERMAN, S.R. and MANSKI, C.F. (1976) "Alternative sampling procedures for calibrating disaggregate choice models", Transportation Research Record 592, pp 24-28.

LERMAN, S.R. and MANSKI, C.F. (1979) "Sample design for discrete choice analysis of travel behaviour: the state of the art", Transportation Research, Vol. 13A, No. 1, pp 29-44.

LERMAN, S.R., MANSKI, C.F. and ATHERTON, T.J. (1976) "Non-random sampling in the calibration of disaggregate choice models", Final Report to the Urban Planning Division, Federal Highway Administration, U.S. Department of Transportation, Washington D.C., Contract FHWA PO-6-3-0021, by Cambridge Systematics Inc.

LIOU, P.S., COHEN, G.S. and HARTGEN, D.T. (1975) "An application of disaggregate mode choice models to systems-levels travel demand forecasting", New York: New York State Department of Transportation, Planning Research Unit, Preliminary Research Report 75.

LOUVIERE, J.J. and MEYER, R.J. (1979) "Behavioural analysis of destination choice: theory and empirical evidence", Iowa City: University of Iowa, Institute of Urban and Regional Research, Tech. Report 112.

MANHEIM, M.L. (1979) "Readings in Transportation Systems Analysis", Cambridge: Massachusetts Institute of Technology, Center for Transportation Studies, Preliminary Edition, August 1979.

MANSKI, C.F. and LERMAN, S.R. (1977) "The estimation of choice probabilities from choice based samples", Econometrica, Vol. 45, No. 8, pp 1977-1988.

MANSKI, C.F. and McFADDEN, D. (1980) "Alternative estimators and sample designs for discrete choice analysis", in C.F. Manski and D. McFadden (Eds.), Structural Analysis of Discrete Data: with Econometric Applications, Cambridge: M.I.T. Press.

McFADDEN, D. (1976) "The theory and practice of disaggregate demand forecasting for various modes of urban transportation", Berkeley: University of California, Institute of Transportation Studies, Working Paper 7623.

McFADDEN, D. (1978a) "Modelling the choice of residential location", in A. Karlqvist, L. Lundqvist, F. Snickars, and J.W. Weibull (Eds.) Spatial Interaction Theory and Planning Models. Amsterdam: North Holland.

McFADDEN, D. (1978b) "On the use of probabilistic choice models in economics", American Economic Association Meeting, Chicago, August 1978.

McFADDEN, D. (1979a) "Econometric models of probabilistic choice", Cambridge: Massachusetts Institute of Technology, Department of Economics, Working Paper (unpublished)

McFADDEN, D. (1979b) "Quantitative methods for analysing travel behaviour of individuals: some recent developments", in D.A. Hensher and P.R. Stopher (Eds.) Behavioural Travel Modelling. London: Croom Helm.

McFADDEN, D. and REID, F. (1975) "Aggregate travel demand forecasting from disaggregate behavioural models", Transportation Research Record 534, pp 24.37.

McFADDEN, D., TYE, W. and TRAIN, K. (1976) "Diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model", Berkeley: University of California, Institute of Transportation Studies, Working Paper 7616.

McINTOSH, P.T. and QUARMBY, D.A. (1970) "Generalised costs and the estimation of movement choice and benefits in transport planning" London: Department of the Environment, M.A.U. Note 179.

MEYBURG, A.H., and STOPHER, P. (1975) "Aggregate and disaggregate travel demand models", Transportation Engineering Journal of ASCE, Vol. 101, No. TE2, pp 237-245.

MILLER, D.R. (1974) "Aggregation problems" Transportation Research Board, Special Report 149, pp 25-30.

ORTUZAR, J.D. (1978) "Mixed-mode demand forecasting techniques: an assessment of current practice", PTRC Summer Annual Meeting, University of Warwick, July 1978.

ORTUZAR, J.D. (1979) "Testing the theoretical accuracy of travel choice models using Monte Carlo simulation", Leeds: University of Leeds, Institute for Transport Studies, Working Paper 125.

ORTUZAR, J.D. (1980a) "Mixed-mode demand forecasting techniques", Transportation Planning and Technology, Vol. 6, No. 2,pp 81-96.

ORTUZAR, J.D. (1980b) "Modelling park 'n' ride and kiss 'n' ride as submodal choices: a comment", Transportation, Vol. 9, No. 3 (in press).

ORTUZAR, J.D. (1980c) "Modal choice modelling with correlated alternatives: Applications to mixed-mode travel demand forecasting", Leeds: University of Leeds, Institute for Transport Studies, Ph.D. Thesis (unpublished)

ORTUZAR, J.D. and WILLIAMS, H.C.W.L. (1978) "A geometric interpretation of random utility models of choice between discrete alternatives", Leeds, University of Leeds, Institute for Transport Studies, Tech. Note 3 (unpublished).

PRASHKER, J.N. (1979) "Mode choice models with perceived reliability measures", Transportation Engineering Journal of ASCE, Vol. 105, No. TE3, pp 251-262.

PROJECTBUREAU INTEGRAL TRAFFIC AND TRANSPORTATION STUDIES (1977) "The SIGMO Study", 4 reports to the Ministry of Transport of the Netherlands, The Hague.

QUARMBY, D.A. (1967) "Choice of travel mode for the journey to work: some findings", Journal of Transport Economics and Policy, Vol. 1, No. 3., pp 1-42.

REID, F.A. (1977) "Roles of behavioural traveller models in policy analysis", Third International Conference on Behavioural Travel Demand Modelling, Australia, April 1977.

REID, F.A. (1978a) "Aggregation methods and tests", Berkeley: University of California, Institute of Transportation Studies, Research Report UCB-ITS-RR-78-6.

REID, F.A. (1978b) "Systematic and efficient methods for minimising error in aggregate predictions from disaggregate models", Transportation Research Record 673, pp 59-124.

RUIJGROK, C.J. (1979) "Disaggregate choice models: an evaluation", in G.R.M. Jansen, P.H.L. Bovy, J.P.J.M. Van Est and F. Le Clercq (Eds.) New Developments in Modelling Travel Demand and Urban Systems, Westmead, Farnborough: Saxon House.

SENIOR, M.L. and WILLIAMS, H.C.W.L. (1977) "Model-based transport policy assessment, 1: the use of alternative forecasting models", Traffic Engineering and Control, Vol. 18, No. 9 pp 402-406.

SHEFFI, Y., HALL, R. and DAGANZO, C. (1980) "On the estimation of the multinomial probit model", International Conference on Research and Applications of Disaggregate Travel Demand Models, University of Leeds, 14-16 July, 1980.

SOBEL, K.L. (1980) "Travel demand forecasting with the nested multinomial logit model", Fifty-ninth Annual Meeting of the Transportation Research Board, Washington D.C. January 1980.

SPEAR, B.D. (1976) "Attitudinal modelling: its role in travel demand forecasting", in P.R. Stopher and A.H. Meyburg (eds.) Behaviour Travel Demand Models. Lexington, Massachusetts: Lexington Books, D.C. Heath and Co.

SPEAR, B.D. (1977) "Applications of new travel demand forecasting techniques to transportation planning: a study of individual choice models", Washington D.C.: Federal Highway Administration, U.S. Department of Transport, Office of Highway Planning

SPEAR, B.D. (1979) "Travel-behaviour research: new directions for relevancy", Fourth International Conference on Behavioural Travel Modelling 1979, Eibsee, July 1979.

STOPHER, P.R. (1975) "Goodness-of-fit measures for probabilistic travel demand models", Transportation, Vol. 4, No. 1, pp 67-83.

STOPHER, P.R., SPEAR, B.D. and SUCHER, P.O. (1974) "Towards the development of measures of convenience for travel modes", Transportation Research Record 527, pp 16-32

STOPHER, P.R. and WILMOT, C.G. (1979) "Work-trip mode-choice models for South Africa", Transportation Engineering Journal of the ASCE, Vol. 105, No. TE6, pp 595-608.

TALVITIE, A.P. and KIRSCHNER, D. (1978) "Specification, transferability and the effect of data outliers in modelling the choice of mode in urban travel", Transportation, Vol. 7, No. 3, pp 311-332.

TARDIFF, T.J. (1976) "A note on goodness-of-fit statistics for probit and logit models", Transportation, Vol. 5, No. 4, pp 377-388.

TRAIN, K.E. (1977) "Valuations of modal attributes in urban travel: questions of non-linearity, non-genericity, and taste variations", Cambridge Systematics Inc. West (unpublished).

TRAIN, K.E. and McFADDEN, D. (1978) "The goods/leisure trade-off and disaggregate work trip mode choice models", Transportation Research, Vol. 12, No. 5, pp 349-353.

WATANATADA, T. and BEN-AKIVA, M.E. (1978) "Spatial aggregation of disaggregate choice models: an areawide urban travel demand sketch planning model", Fifty-seventh Annual Meeting of the Transportation Research Board, Washington D.C., January 1978.

WERMUTH, M.J. (1978) "Structure and calibration of a behavioural and attitudinal binary mode choice model between public transport and private car", PTRC Summer Annual Meeting, University of Warwick, July 1978.

WILLIAMS, H.C.W.L. (1977) "On the formation of travel demand models and economic evaluation measures of user benefit", Environment and Planning A, Vol. 9, pp 285-344.

WILLIAMS, H.C.W.L. (1979) "Travel demand forecasting: an overview of theoretical developments", Research Seminar on Transport and Public Policy Planning, University of Reading, March 1979.

WILLIAMS, H.C.W.L. and ORTUZAR, J.D. (1980a) "Behavioural theories of dispersion and the misspecification of travel demand models", Working Paper 277, School of Geography, University of Leeds.

WILLIAMS, H.C.W.L. and ORTUZAR, J.D. (1980b) "Travel demand and response analysis - some integrating themes", International Conference on Research and Applications of Disaggregate Travel Demand Models, University of Leeds, 14-16 July, 1980.

WILLIAMS, H.C.W.L. and SENIOR, M.L. (1977) "Model based transport policy assessment, 2: removing fundamental inconsistencies from the models", Traffic Engineering and Control, Vol. 18, No. 10. pp 464-469.