choices actually made and observed. Without prejudice to the ultimate question of whether stated preference data can be legitimately used, it is generally accepted that it has a greater chance of reliability if the circumstances of the hypothetical choice are reasonably within the experience of the respondent. Most studies hence make some attempt to provide an appropriate context for hypothetical questions. We will discuss this in more detail later.

The dichotomy between the market research emphasis and that of transport modellers is basically that market researchers have concentrated on survey techniques, while transport modellers have increasingly concentrated on statistical theory. The result is considerable confusion over nomenclature. From the transport modeller's point of view, several techniques which are distinguished by market researchers (primarily on account of their differing survey techniques) appear to show no substantive difference in their model assumptions.

Although it is proper to bear in mind the connection between data collection and analysis, it is necessary to clarify the process by which techniques are distinguished, and a logical step is to classify techniques both by their data requirements and by their model assumptions. Most transport planners have much to learn from a more thorough understanding and appreciation of the survey methods developed by market researchers, whereas in general market researchers would profit enormously from a better understanding of the statistical underpinnings of their models.

Having said that, this paper will concentrate on the model assumptions rather than the survey techniques, noting interdependence where crucial. To aid our discussion, we will try to standardize the concept of model.

Although there are a large number of models of individual choice, particularly within the field of psychology, the most well-known choice models are those derived from the concept of utility, and we will refer to them as 'random utility models'; included in this group is the multinomial logit model (MNL), which has widespread popularity because of its flexibility and relative simplicity. The basic notion of a random utility model is that for each alternative in the choice set it should be possible to calculate the utility corresponding to that alternative, as the sum of a deterministic element and a random element. The deterministic element typically contains information about the attributes of the alternative, weighted by suitable coefficients (which are normally estimated by statistical means), while the random element may deal with the effect of unidentifiable or unobservable variables, general "noise", etc. The respondent is assumed to choose that alternative which offers him the highest utility.

While these kind of models have tended to be calibrated on revealed preference data, a number of well-known problems have been encountered. With revealed preference data, we know which alternative was chosen, and with this knowledge we may proceed to measure certain of its attributes more or less independently of the respondent. However, there may be difficulties in finding out what alternatives were considered by the respondent, and, inasfar as the respondent is asked for details of the attributes of his rejected alternatives, these details may be very far removed from reality.

To put the problem in the context of the well-known modal split model, let us assume that we have been able to leave aside the question of choice set definition, and the reliability of reported attributes, and see what is involved in calibrating a choice model. For ease of illustration we shall assume that only two variables (cost and time) enter into the utility function, but the example can readily be extended to more variables. In such a case, the difference in utility for the two alternatives (say car or train) can be written as:

$$DU = a + b\ DC + c\ DT + De$$

assuming a linear formulation; here e is the random element.

Now the equation $0 = a + b\ DC + c\ DT$ represents a line in the (DC,DT) plane, as shown in Figure 1 a, the actual location and slope of the line being determined by the (unknown) values of a,b and c. If De is small, then it will approximately be the case that any observation on one side of this line will choose one alternative (say, car), while any observation on the other side will choose train. Thus, the modelling process can be seen as one of choosing a line which will as accurately as possible segment the population into those who choose car and those who choose bus.

Consider now the data illustrated in Figure 1 b. Here the observations have been plotted according to their values of DC and DT, and have also been coded according to their choice. On the basis of what has just been said, it will be appreciated that the data provides virtually no help in locating the line segmenting the population into car and train choosers. Each respondent has chosen the "dominant" alternative - ie, the alternative which is favoured on all the attributes. In such a case, it will be impossible to define the coefficients a,b and c with any reliability.

The situation is not much better in Figure 1 c, although at least in this case there is some evidence of a possible tradeoff between DT and DC. It may thus be appreciated that what is required in order to calibrate a satisfactory model is not to have clearly separated population groups with distinct

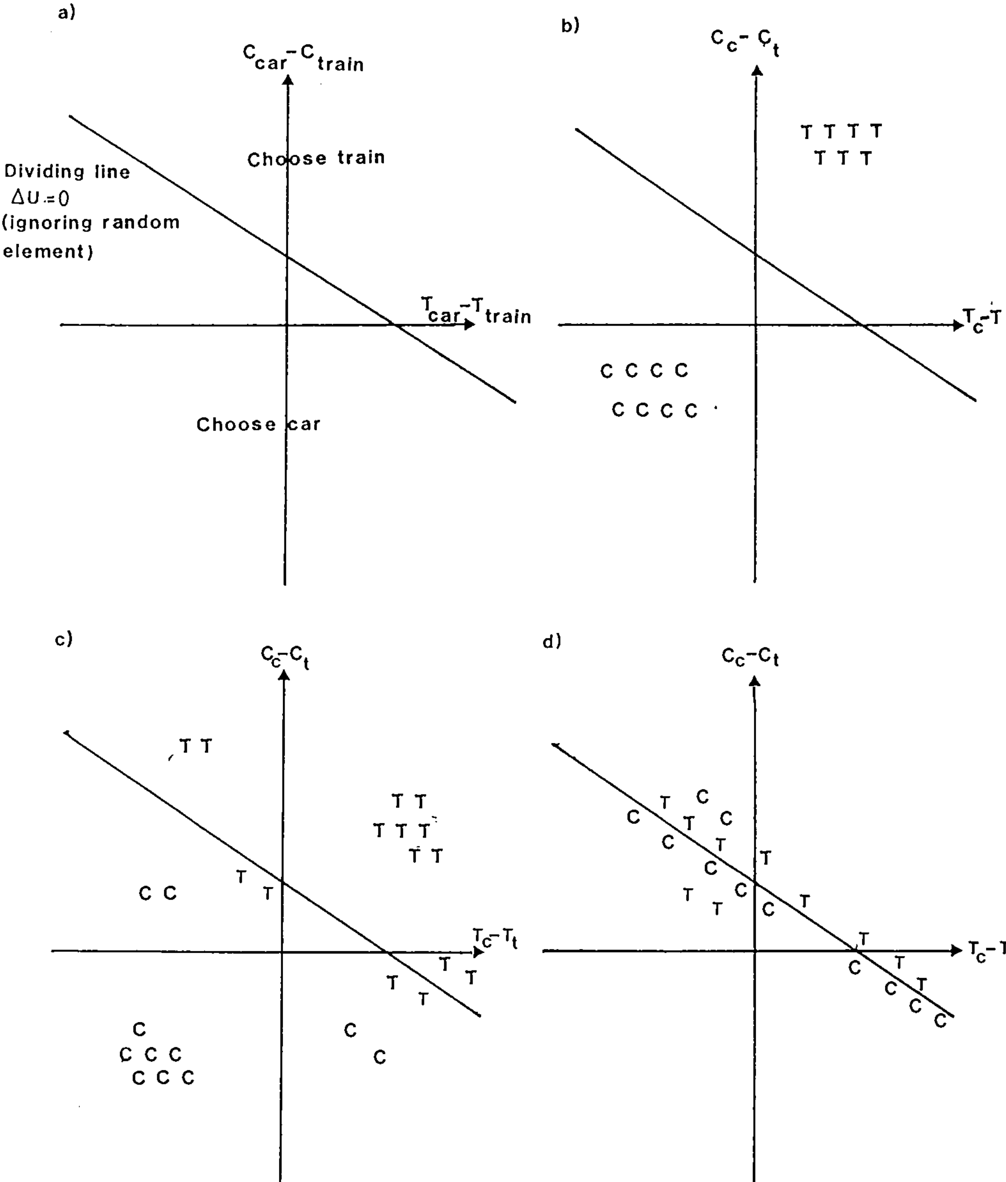

Figure 1 Hypothetical illustration of mode choice data

choices, but to have as many "marginals" as possible - ie respondents who might choose a different alternative given a small change in the attributes DT and DC. As can be seen in Figure 1 d, this involves having as many respondents as possible located adjacent to the line of discrimination.

An additional requirement well-known to model-builders is that the variables in the equations (DT and DC in this case) should themselves not be too closely correlated, otherwise it will not be possible to identify separate effects for them.

All in all, the requirements of data for the calibration of choice models based on revealed preference are quite exacting, and many of these requirements are not at the control of the model-builder. Much of the data which is collected may be of very little help in actually calibrating the model, even if the survey is well designed. Consequently, sample sizes may need to be increased to achieve tolerable accuracy, and this may have serious cost implications.

The market research approach is very different (A useful survey of market research techniques is provided in Green and Srinivasan (1978)although it is no longer completely up to date). Most methods rely on an experimental design, such as fractional factorials or Latin squares; this allows the researcher much more control over the structure of his data, and the attributes can be clearly specified. For instance, there might be three attributes influencing choice, each being presented at three levels; hence 27 different alternatives. A fractional factorial design would provide an efficient way of reducing the number of alternatives to something more manageable.

The respondent is normally asked either to assess different alternatives (using various "rating scales", which may be defined verbally - eg "very good", "satisfactory" etc. - or "interval scales" - eg 1,2,3,4,5), or to rank a set of alternatives in order of preference.

The various techniques which have been proposed differ principally in the way in which they organize the tasks which the respondent has to carry out, in the interests of improving the quality of the data. For instance, one of the variants, known as "trade-off analysis" presents the respondent with a sequence of ranking tasks on subsets of the total range of possibilities: within each subset, only two of the attributes are altered. It is claimed that in this way, the task of the respondent is made easier and hence it is hoped that the data will be more reliable. Trade-off analysis has been used in a number of studies carried out for the New York State Department of Transport: see for example Koeppel(1977) and Eberts & Koeppel (1977).

It has been found (for instance, by Eberts & Koeppel (op.cit.) that that problems may be caused by "respondent fatigue" - the willingness and ability of the respondent to answer questions may decline after a time. For this and other reasons, it is normal practice to randomize the order in which tasks or alternatives are presented. It is also generally accepted as important to provide a context of realism within which the alternatives can be assessed.

As far as the analysis of the market research data is concerned, the general practice until recently has been to use fairly crude methods. For instance, most of the applications of conjoint analysis appear to have assumed that ranked alternatives can be located at equal intervals along some preference scale - not a very appealing assumption. It should be said however that because of the general lack of emphasis on the precise details of the model and software used, it is quite difficult to deduce from published work exactly what methods have been used.

Given the generality of the choice model, it is equally suitably applied to revealed preference data as to stated preference data. Although it is only very recently that such models have been applied to stated preference data within the area of market research, much of the previous analysis has been based on essentially similar concepts, but with less theoretical rigour. The application of the body of knowledge relating to discrete choice models to market research data can be seen as a major advance in terms of statistical content.The fact that the same model can be applied to both kinds of data makes comparative work a real possibility.

3. A BRIEF DISCUSSION ON RANKED DATA

As noted above, the data obtained from stated preference experiments usually consists either of an explicit rating for each option, or a ranking. In the case of explicit ratings, the data can be treated as utility or probability scores (after appropriate transformations, where required) and analysed by standard multi-variate analysis techniques (multi-linear regression, analysis of variance etc.). Rankings can best be analysed within the framework of discrete choice analysis by interpreting the data as being the choices made from successively limited choice sets.

This approach has been used by, for example, Punj & Staelin (1978), Chapman & Staelin (1982), and, specifically within the transport field, by Beggs, Cardell & Hausman(1981). Without loss of generality, if there are n alternatives which are ranked by an individual in the order 1,2,....n, then this can be interpreted as a series of (n-1) subchoices whereby alternative 1 is preferred out of the whole set, then alternative 2 is preferred out of the whole set excluding alternative 1, and so on.

Within the framework of random utility models, it is relatively simple to write down the likelihood function of the set of ranked alternatives (see Beggs, Cardell & Hausman(op.cit.) eq. 2); the problem consists in carrying out the computation, on account of the multiple integrations involved. However, if the random element in the model can be assumed consistent with the multinomial logit formulation, then two great simplifications are made. Firstly, the need for multiple integration is removed, since the multinomial formulation allows the integrals to be reduced analytically; secondly, according to Chapman & Staelin, the series of (n-1) subchoices can be treated as independent observations. This allows the data set to be decomposed into a much larger number of observations, which can be analysed using standard multinomial logit software.

Two caveats needs to be mentioned here. The first is the well-known limitation of the multinomial logit model, that the error terms attached to each alternative's utility function should be independently and identically distributed. Although there are ways of alleviating the effects of this assumption, it remains a potentially serious restriction. However, against this it can be said that in analysing ranked data in the way suggested here, the assumptions have been made quite clear; not only does this represent a significant improvement over earlier market research work, but the particular assumptions that are made open the way for a comparison between stated preference and revealed preference methods.

Secondly, there is a reasonable likelihood that the quality of ranked data may not be consistent throughout the set of alternatives. Although not much is known about the process whereby individuals actually go about the task of ranking a set of alternatives, it seems plausible that the ranking among the less preferred alternatives may be less reliable than that among the more preferred alternatives. Along these lines, Chapman & Staelin suggest that it is not necessarily worth decomposing the data into the full (n-1) separate decisions, and propose some criteria to assist with judging how far to decompose the data.

4. CALIBRATION AT THE INDIVIDUAL LEVEL

Current models of discrete choice calibrated on revealed preference data are often termed disaggregate, in recognition of the fact that data is available at the individual level. It is however inconceivable to calibrate such models for each individual separately, because of the shortage of information on individual choice patterns. In calibrating revealed preference models, therefore, some assumption has to be made about the consistency of the postulated utility function over the members of the population; this is the well-known problem of taste variation. A simple way of dealing with this potential problem is to define relevant market segments, and to calibrate the model separately to each segment.

In the case of stated preference models, it is in fact often possible to calibrate the model separately for each individual, although with not great precision. This allows a different approach to the market segmentation, since this can be done by a comparison of the coefficients for individuals. Individuals with sets of coefficients which can be judged statistically similar can be grouped into segments, and the model can then be re-estimated at the segment level to improve the precision of the coefficients. The usual likelihood ratio test is available to allow us to judge the suitability of the segmentation.

Such an approach, to the best of the author's knowledge, has not been reported in the literature, although one of the versions of conjoint analysis estimates coefficients for each individual and then averages the coefficients in a very simple way. This probably stems from a failure to appreciate the statistical theory of choice models, which allows a comparison of the individual coefficients to be made. A rigorous treatment of the method described in the previous paragraph could do much to resolve the problems relating to taste variations.

5. COMPARISON OF THE TWO SOURCES OF DATA

In a recent article by Louviere et al (1981), it is pointed out that the revealed preference and stated preference approaches are to a considerable extent complementary. The revealed preference approach has the major advantage that it is related to observed data. However, as has already been pointed out, this advantage is considerably diluted by the difficulty of defining the choice set, concern about the accuracy of the data actually used in making the choice, and a lack of a priori information about the accuracy with which coefficients can be estimated (apart from the experience gained in fitting comparable models to other sets of data).

All these disadvantages are resolved with the stated preference approach, but the crucial question to be asked is whether the answers given by respondents relate in any way to the decisions that they would make in practice. Something is known of circumstances in which the answers can be expected to be invalid: we have already referred to response fatigue, and a low response rates or a refusal to carry out certain tasks, is another indication. There are also various tests relating to consistency, and the "randomness" with which responses are made. While a careful procedure with respect to these factors will eliminate the worst failings of the approach, there is still no guarantee that the results will be reasonable.

(We may also notice in passing that both kinds of data will produce response problems, in that the sample of respondents will differ from the original sample base not only in size but very possibly in terms of representativeness (response bias). Very little is known about the different response rates likely to be related to the two types of data).

This discussion suggests that there are two kinds of tests which need to be performed in order to assess the value of stated preference analysis. The first is on the predictive ability of both stated preference and revealed preference methods; the second is a controlled comparison of the actual models produced by the two methods.

The question of predictive ability is a somewhat thorny one, since it relates to at least two issues: the transferability (over time) of the model, and the changes over time in the input data. Although theoretically these can easily be distinguished, in practice it is often very difficult to disentangle effects, and to say with confidence that a model's failure to predict a certain outturn is due to model specification rather than an incorrect prediction of the input variables. In addition, models are often used to predict "new situations" (eg the introduction of a new mode), where there is a danger of extrapolating outside the reasonable range of the model. Little work has been done on transferability over time; rather more has been done on spatial transferability for revealed preference models and the results can only be described as mixed. Certainly there is no evidence of global consistency. It would seem fair to conclude that any claim for the longer term predictive ability of revealed preference models remains unproven, but that some short term predictions have appeared to be satisfactory.

A number of claims have been made for the predictive ability of stated preference models, but since most of the work has been done in the field of market research, the details are usually unavailable on grounds of commercial confidentiality. Applications are typically restricted to short term forecasts. We note, however, that Louviere et al (1981) refer to "consistent evidence amassed over the past five years that models built on responses to hypothetical scenarios are accurate predictors of real behaviour in analogous situations".

Although some information about reliability may be gained from an examination of predictive accuracy, it appears that a more convincing way to increase our confidence in the modelling process in general, and stated preference data in particular, is to compare the models calibrated on data relating to the same individuals. For this purpose, we require a data set large enough to allow a satisfactory revealed preference model to be fitted, and that the respondents should also have been asked stated preference questions.

Recent work by Louviere et al (1981) attempts to carry out such a comparison. Although the details are not completely clear, it appears that nearly 800 usable questionnaires were obtained, with data relating to mode choice. However, these were divided between two different towns and two points in time; the first point in time had 263 usable questionnaires and the second 516, but no information is given about the split between towns. All respondents provided data on both stated preference and revealed preference questions, and for the stated preference data, 30 separate scenarios were offered.

Louviere et al begin by fitting models to the stated preference data, separately for the four surveys. The stated preference data is based on ten attributes, and most of these variables are entered both linearly and quadratically. Because the variance in the independent variables is controlled, the accuracy with which the coefficients can be determined is effectively fixed. It is concluded that the data can be merged across towns but not across points in time.

Next, a similar model is calibrated on the revealed preference data. It appears that the response variable is in fact a measure of relative frequency with a logit transform, rather than the (0,1) variable of most discrete choice models, and this should improve the accuracy of estimation. But it turns out that the level of accuracy is poor; very few of the coefficients are significantly different from zero. It is hypothesized that part of the problem may be due to correlation between the independent variables and personal factors; consequently, some twelve personal variables were added to the model, and the data was aggregated across towns.

No measures of goodness-of-fit are given for this combined model, but it may be noted that in the most favourable case (the second point in time) there are 516 observations with 25 variables, of which, depending on the response variable used, only five or ten have significant coefficients even at the 10% level (at the 5% level, the numbers are correspondingly two and six). With such low levels of significance, it is not surprising that for almost all the coefficients, it is impossible to reject the hypothesis that the coefficients resulting from the stated preference and revealed preference models are the same!

It is worth dwelling on this piece of work, for two reasons: in the first place, it is to the author's knowledge the only piece of comparative work that has been published, and secondly, it highlights a number of the problems that may be encountered in such work. Let us consider the nature of these problems.

In the first place, it is essential that a reasonably successful revealed preference model can in fact be calibrated: this does not seem to have been the case with Louviere et al. For reasons given in Section 2 of this paper, this is likely to involve careful survey design, aimed at obtaining a sufficient number of respondents for whom the chosen mode is not dominant, and at the same time ensuring that the independent variables are not too highly correlated. Any comparison between the two methods will require a satisfactory level of accuracy for the coefficients of the revealed preference model.

Secondly, attention must be paid to whether the coefficients should in fact be the same, as opposed to, say, merely having the same relative values. This question basically relates to the statistical assumptions underlying the model. Since the utility formulation can only be determined up to a monotonic transformation, some assumption is necessary to obtain a determinate solution, and these assumptions are not always made explicit. For instance, in the standard multinomial logit model, the coefficients are scaled relative to the standard deviation of the random element, which is fixed at a constant value by assumption. Care would thus be needed in reconciling such coefficients with, say, a multiple regression analysis on a logit transform of a continuous response variable.

In the example given by Louviere et al, this problem is avoided by using the same dependent variable (frequency of mode choice) for both sets of data. Of course, the very choice of this variable in the revealed preference case may present some difficulties, in that it is more likely to be subject to reporting errors than the usual "yesterday's mode" question. However, the comparison will clearly be simplified if the response variables for both types of data are the same. There remains a need to clarify the relationship between models which have a common utility formulation but a different form of the response variable.

Thirdly, there is the crucial question of what kind of statistical comparison should be made between the two sets of coefficients. This appears to be currently unresolved, but it does not seem reasonable to treat them as independently derived estimates, given that they are obtained from identical samples, and are intended to relate to the same decision process. A secondary question is whether it is sufficient to carry out pairwise comparisons on corresponding coefficients, or whether some more global measure should be used, which takes account of the covariance within the set of estimated coefficients.

It will be noticed that we are not making any claims within this paper as to whether stated preference models are intrinsically better or worse than revealed preference models. However, the reality of the situation is that within the transport field, revealed preference models have achieved a considerable level of acceptance, despite scepticism from some quarters. Thus, regardless of the hypothesized merits or demerits of either type of model, it seems that, practically speaking, increased acceptance of stated preference models will depend on their ability to achieve compatibility with revealed preference models.

6. PROPOSALS

What is required, with some urgency, is a number of reliable tests comparing the two types of models. The essential component for this - apart from the solution of some of the statistical questions referred to above - is a well-designed revealed preference survey which includes stated preference questions. The simplest way to achieve this is to insert stated preference questions into revealed preference studies that are already being funded and carried out.

The author is currently involved in three such studies in collaboration with Martin & Voorhees Associates. One study is concerned with long distance travel in the Netherlands, another relates to mode choice in the West Midlands conurbation of England, while the third is in connection with a study to measure the value of travel time savings in various contexts. Results from these studies will be available in due course.

The additional cost imposed on the "parent" study by tagging on a number of hypothetical questions is virtually zero. The most persistent concern - that the difficulty of dealing with such questions might prejudice response overall - does not seem to be justified. In fact, when stated preference data has been collected on its own, surprisingly high response rates have been obtained, even with postal questionnaires.

In this way, the necessary stated preference data can be collected virtually for free, apart from the cost of the experimental design, since the parent study is committed to the cost of carrying out the survey, and indeed of building the revealed preference model. Given these considerable advantages, the main concern of the analyst carrying out the comparison is that the revealed preference model has a chance of being successfully calibrated, and, as discussed earlier, this question relates principally to survey design.

It would thus be extremely useful to prepare a list of studies which are currently under consideration where it is intended to fit random utility choice models to data relating to choices actually made, and on the basis of such a list, decide which studies offer suitable opportunities for a comparative exercise along the lines suggested in this paper.

7. CONCLUSION

The growing convergence between traditional econometric techniques and those of market research has led to an increased interest in data collection methods, while at the same time strengthening the theoretical basis of market research analysis. Given the large potential advantages of using such techniques within the transport context, it is important to validate people's ability to respond consistently to hypothetical choice questions.

Since the revealed preference approach is widely accepted by transport modellers, the best course of action is to find as many such studies as are currently under consideration as possible, and tag on suitable stated preference questions, so that models can be calibrated using both kinds of data for the same set of individuals.

A brief discussion of one comparative study carried out by Louviere et al revealed a number of problems that need to be solved. The most important is to ensure that a satisfactory revealed preference model can be calibrated. Next, any possible reasons for finding different coefficients that relate to the model specification need to be clarified. Finally, the basis of the statistical tests for comparing the two models requires some elaboration.

If all these problems can be solved, and a number of well-conducted comparative studies are carried out, there are two potential outcomes. Either the stated preference models will be found, on balance, to be compatible with revealed preference models, in which case there should be no argument about a much greater use of stated preference techniques, or they will be found to be incompatible. If the latter is true, then the validity of either technique can only be established against the criterion of predictive ability. This criterion should of course be the basis for preferring any type of model over another. However, as has been pointed out in this paper, investigations of predictive ability encounter considerable problems. If this turns out to be the only way of adjudicating between the two approaches, it is likely that the current controversy will continue for some time to come.

8. ACKNOWLEDGEMENT

I should like to thank Mick Roberts of Martin & Voorhees Associates for his helpful coments on an earlier draft. In addition, I acknowledge my indebtedness to Hugh Gunn, now of Cambridge Systematics Europe, from whom I have learnt much in collaboration over a number of years. He has not, however, been involved in preparing this paper, and bears no responsibility for any errors or misconceptions in it.

9. REFERENCES

S Beggs , N S Cardell & J Hausman (1981), Assessing the potential demand for Electric cars, Journal of Econometrics (16), pp 1-19.

R G Chapman & R Staelin (1982), Exploiting rank ordered choice set data within the stochastic utility model, Journal of Marketing Research vol XIX, pp 288-301

P M Eberts & K-W P Koeppel (1977), The trade-off model: empirical and structure findings, New York State Department of Transport (NYSDOT), Planning Research Unit, Preliminary Research Report 123.