2. Forecasting behaviour: with applications to transport

Andrew Daly

INTRODUCTION

Choice modelling has developed enormously over the last few decades as the principal methodology for the quantitative understanding of human behaviour. In particular choice modelling has been used intensively for the determination of people's willingness to pay (WTP) for goods and aspects of goods, both in market and non-market contexts. The data bases for this work have been dominated by the collection of stated preferences and in particular stated choices. Choice modelling applications now cover a wide range of economic sectors: transport, health, environmental analysis, marketing and other fields.

These developments and the expansion of the field can only be viewed as success. However, it is interesting to contrast the objectives and basis of this more recent work with the early development of choice modelling, which began in earnest in the 1960s. At that time, the focus of the work was on *forecasting behaviour*, based on observations of *revealed preferences* (i.e. what people are observed to do) and largely focussed on transport applications. While not questioning the recent success of choice modelling, it is interesting to look at the way in which choice modelling has developed away from its earlier methods and applications and to consider whether there is anything for modern choice modelling to learn from those original areas and methods. I shall argue that we can improve our work and its applicability by revisiting the linked issues of forecasting and the use of revealed preference data.

The role of transport applications in the early development of choice modelling was crucial. There were good reasons for this. First, the important issues of forecasting the patronage and appraising the benefit for new transport infrastructure generated funds, because choice modelling was able to offer more plausible solutions to the problems faced in transport planning than the methods used previously. Second, the nature of the data available for transport planning, where the characteristics of the

alternatives vary widely because different travellers make different journeys, introducing variation over the different origins and destinations of those journeys, meant that even when relatively simple revealed preference data was used it was possible to develop interesting models. Third, numerical analysis is familiar to transport analysts, who often have quantitative backgrounds. The fruitful symbiosis of transport analysis and choice modelling in its early years is discussed in the first section of the chapter.

Referring to this early research, I argue in the second section that the analysis of revealed preference data remains fruitful, for a number of reasons, in particular that it gives a much better basis for forecasting. Stated choices do not equate to actual behaviour. Although planners may want people to use buses or cycle, and people may state that they would choose those alternatives in the comfort of an experiment in their home or a laboratory setting, the reality of cold, rain and the unreliability of the service may reveal a preference for car travel.

In its turn, forecasting raises a number of issues, discussed in Section 3, that have not received much attention in the choice modelling literature. These issues are difficult and important, so that they would form suitable subjects for academic study. Reliable forecasts are essential for many important policy issues, such as decisions on transport infrastructure or new product launches, so that improving and estimating the accuracy of forecasts has clear societal benefit.

A final section of the chapter suggests lines for research that seem to be promising.

1. CHOICE MODELLING IN THE TRANSPORT FIELD

The origins of practical choice modelling owe much to the efforts of workers in the transport sector. Initially, it seems most progress was made in the UK, but subsequently American workers took over the leadership of the field. The steps taken by these workers and in particular the reasons for their decisions give a lot of insight into the early development of the field, while leaving open lines of research that were closed then but can be reconsidered now in the new context of the 21st century.

1.1 British work in the 1960s

The history of choice modelling with transport applications is long and distinguished. One can see a choice-modelling basis in the 'principle' enunciated by the British highway engineer J.G. Wardrop (1952), still the

explicit basis for important components of travel demand forecasting 60 years later:

The journey times in all routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route.

Wardrop's principle recognises that drivers have a choice and select routes on the basis of their characteristics (journey times).

The practice of transport planning developed powerfully in the UK in the following years and by the end of the 1960s it seems clear that British analysts were leading a world that was far less integrated than it is today, when information technology, widespread use of English and cheap travel make it straightforward to keep abreast of the latest developments. Key developments in the UK in that period, which began to lay a foundation for transport analysis and choice modelling more generally, included the following.

- The work of Beesley (1965) aimed at obtaining the 'value of time', interpreted as travellers' WTP for a minute of time savings. Crucially, he saw the appropriate way to approach the problem as being to observe the behaviour of travellers and to interpret their choices as revealing preferences for one combination of time and cost over another. This is essentially the basis used in modern willingness-to-pay studies, usually with the replacement of revealed preference by stated preference information. Beesley was not an econometrist and his estimation procedure was simply to find the value of time that was consistent with the maximum possible number of observed choices; interestingly, this 'score maximisation' method was later shown by Manski (1975) to be a consistent estimator of the true value of time (if such exists) under quite general conditions. Because of the computational challenges which affected all the early work, Beesley used a graphical method, which became known as the Beesley graph.
- The original intention of Quarmby's (1967) work was attempting to predict travellers' mode choice. His contribution was to see that mode choice was not based solely on time and cost but needed to take account of other aspects of choice. Some of these, like the distribution of travel time across walking, waiting and 'in-vehicle' travelling time, could be measured but others, like the general convenience of a car, could not and had to be represented by an estimated constant. Thus the notion of a multi-dimensional choice criterion was established and the linear form, still used in almost all

choice models, was defined. Quarmby also suffered from a lack of software but a computer was essential to process his data in multiple dimensions: he settled on discriminant analysis, an approach that can easily be criticised but in the context of the mid 1960s may well have been the best method available. Referring to Beesley, Quarmby also realised that the coefficient ratios coming out of his discriminant analyses could be interpreted as trade-off ratios, including the money value of time.

- McIntosh and Quarmby (1970) set out the basis for travel demand modelling and transport project appraisal that is still in use today. Using Beesley's and Quarmby's earlier work, they defined the 'generalised cost' to comprise the monetary cost and the time components, converted to monetary equivalents by multiplying by the value of time, which might be different for different time components. Not only could this concept be used in predicting demand, but it could also form the basis for an economic appraisal of transport projects, using the 'rule of a half' to measure the consumer surplus and recognising that time savings were usually the main benefit of transport projects so that their quantification in monetary terms was essential to a proper appraisal of transport policy proposals.

The key characteristic of this work was that it recognised that travel demand arose from travellers' choices; moreover, these choices could be understood and predicted – with error! – using relatively straightforward and intuitively plausible methods. The specifically British motivation for the work was the need to make relatively rigorous economic appraisals of transport projects and the consequent requirement to make forecasts of demand to feed into the appraisal. These characteristics of UK transport planning were very helpful in forcing British researchers to develop choice modelling on a more rigorous quantitative basis and – importantly – in motivating funding for the research that needed to be done, which represented a small fraction of the infrastructure budgets being considered. In particular, it was necessary to have good estimates of values of time.

The British travel demand work contrasts with that of contemporary studies by the major US engineering companies, which, although carefully done and involving large data collection exercises, gave no scope to any notion of choice. The criticism of their analysis of travel flows as being purely based on physical analogies with only superficial resemblance to the true object of study, i.e. travelling people, was heard quite frequently. While there was some transport choice modelling work in US universities in the 1960s (e.g. Warner, 1962; Lisco, 1967; Lave, 1969), it seems not as systematic or as comprehensive as the contemporary British work, nor

does it seem to have had an impact on practical planning; this was to change in the following decade.

The accusation of the arbitrary use of physical analogy also affects the use of entropy-maximising models, an approach notably followed in the UK by Wilson (e.g. 1967, 1969), who applied these models, developed in physical sciences, in the context of research in geography. Wilson's work, extensive, detailed and internally consistent, was influential on British travel demand forecasting and was the basis for the important SELNEC model (south-east Lancashire, north-east Cheshire, i.e. Greater Manchester). The criticism of not representing real human behaviour has prevented much serious subsequent use of entropy models and reduced the influence of Wilson's work (but see Vrtic et al., 2007). However, it was shown by Miyagi (1984) that entropy maximisation is the dual of cost minimisation in an optimisation sense and therefore can be seen as in principle entirely consistent with the concept of utility maximisation that underlies most modern choice modelling. All of the entropy work can then be reinterpreted as models of choices made by utility-maximising travellers.

Confirming the consistency of entropy with utility modelling, we find in Wilson's 1967 paper, for example, a very early example of the use of a logsum 'composite cost', discussing the formula for the number of trips T_{ij} between zones i and j:

> Note further that the T_{ij} derived in [the entropy model] above can be wholly identified with the T_{ij} derived in the conventional gravity model . . . provided that
>
> $$\exp(-\beta c_{ij}) = \sum_k \exp(-\beta c_{ij}^k) \tag{64}$$
>
> This equation is of the greatest importance because it shows how a composite measure of impedance, $\exp(-\beta c_{ij})$, or average generalized cost, c_{ij}, can be derived from the modal impedances $\exp(-\beta c_{ij}^k)$ where these are known individually. Such composite impedances are valuable in a variety of planning models, but past practice has been to use one of a number of arbitrary averaging procedures.

The focus of UK research work in the late 1960s was in the Mathematical Advisory Unit (MAU), based in the Ministry of Transport, which from 1966 to 1971 produced numerous important pieces of work. However, and disastrously for the quality of British transport analysis, MAU was then wound up. The methods used in travel demand forecasting, even for important studies, reverted to the conventional approaches of the engineering companies. The culmination of this process was the Regional Highway Traffic Model (Alastair Dick & Associates, 1978),

which attempted to develop a model at national scale using conventional approaches but failed expensively in its central aim, though a number of useful by-products were developed. The result was to discredit large-scale travel demand modelling in the UK for 20 years, although smaller models continued to be developed for the estimation of values of time and other specialised studies.

A recurrent theme in the 1960s work is the use of methods that were attacked at the time for not conforming to accepted criteria but which were subsequently shown to conform to those criteria. Manski (1975) showed that score maximisation estimation gave statistically consistent estimations; Miyagi (1984) showed that entropy methods were the dual of utility maximisation. Similarly, discriminant analysis was rejected because it represented the continuous ("explanatory") variables as being dependent on the discrete ("dependent") variable; however, a more modern view would recognise that causality is not so simply assigned and there may be merit in considering a different view. There may also be merit in other approaches that have been discarded on technical grounds, because subsequent work may reinterpret the technical findings.

1.2 The 1970s: US researchers take the lead

US work in transport analysis seems to have been well known in the UK, but all of the British work was done without knowledge of the theoretical developments in choice modelling such as the work of Block and Marschak (1960), Luce (1959, and the retrospective in 1977) or, much earlier, Thurstone (1927). However, it was not until the 1970s that US choice modelling seems to have moved out of the academic context and into practical application, again notably in the transport sector.

The key researcher in this development was, of course, Daniel McFadden. He described (2000) his own initial work in the field as starting in the late 1960s to support a student studying government investment decisions (in the *transport* sector) and continuing after 1970 in a practical travel demand modelling study (Domencich and McFadden, 1975). The policy demands of the transport sector (and its funding) gave impetus to the development of a comprehensive choice modelling framework, based on Random Utility Models (RUM) and the multinomial logit model.

The RUM framework was applied in the key study of the proposed Bay Area Rapid Transit system (BART; see McFadden, 1978a). This showed that RUM methods were capable of making forecasts of the patronage of new modes of transport more accurately and based on a more credible representation of behaviour than the conventional methods. While good fortune may have played a role, as the confidence limits of McFadden's

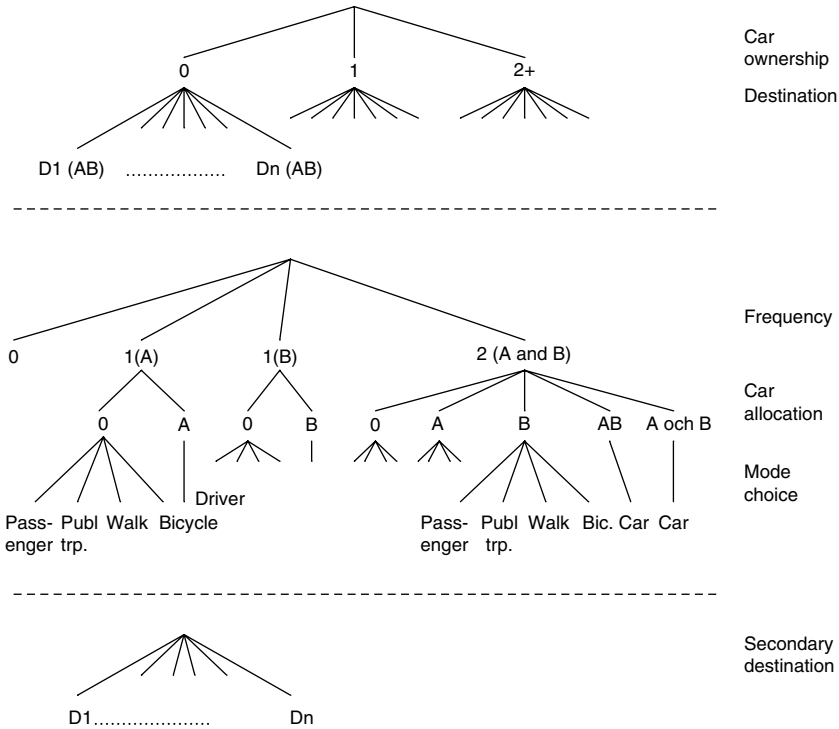
forecasts were actually quite wide, this study was the justification for the gradually increasing adoption of the RUM approach for travel demand forecasting in a number of countries, of which The Netherlands was an early example. The work of Moshe Ben-Akiva and his colleagues at MIT was important in developing, disseminating and applying these ideas. Analysts in other economic sectors could also refer to this success to find methods for their own studies.

RUM analyses were also developed in the UK (e.g. Daly and Zachary, 1975), where the RUM concept was used to integrate forecasting methods with WTP estimations (of values of travel time) and hence with project appraisal. However, it was in and from the US that the largest-scale applications developed.

An outstanding issue in RUM modelling that was also resolved within the transport sector was to establish a theoretical basis for nested logit modelling. The 1975 conference of the International Association for Travel Behaviour noted the problem that nested logit models did not have a well-established basis in RUM theory; by time of the 1977 conference of that organisation there were four independent solutions, later published as McFadden (1978b), Williams (1977), Daly and Zachary (1978) and Ben-Akiva and Lerman (1979). Perhaps these solutions would have been obtained earlier had the authors been familiar with the work of Wilson. The nested models were particularly necessary to meet the requirement in travel demand modelling of dealing with different types of simultaneous choice: choice of travel mode together with choice of destination, route, travel frequency, car ownership, etc. Figure 2.1 shows a complex structure used in a practical model developed for Stockholm a little later, describing the journey-to-work behaviour of couples and illustrating the series of decisions that are taken before a trip is made. Mixed discrete-continuous models could also be developed within the same framework to model (for example) car ownership and use (Train, 1986; de Jong, 1990). Once again the specific needs of the transport sector drove a development which had very wide implications.

Thus by 1980 a basis had been established and a small cadre of researchers was available to take forward RUM modelling to forecast travel demand. The work was based exclusively on revealed preference data, generally collected by home interviews, while other data collection methods were still viewed with suspicion. The unique aspect of travel demand, that a variation in the 'products' available to different people are often substantially different, because of the variation of origins and destinations, made it possible to develop detailed models based on revealed preference data with very limited responses from each individual.

With the election results in the UK in 1979 and in the US in 1980 gov-



Source: Algers et al., 1996.

Figure 2.1 Choice structure for Stockholm commuter model

ernment support for transport projects, and consequently for transport planning research, declined dramatically. From that time, the development of choice models could no longer be so clearly led by the transport sector, but fortunately other sectors began to take up RUM-based choice modelling. Moreover, US/UK research was no longer conducted in isolation from other countries, as international contacts increased. Analysis was no longer based exclusively on revealed preference data. However, the RUM framework and the analysis methods based on it were established and were taken over for further exploitation in other contexts.

In summary, transport problems inspired and funded a substantial part of the early development of choice modelling, which in turn gave better solutions to many of the problems of transport analysis.

2. THE EXPLOITATION OF REVEALED PREFERENCE DATA

The development after about 1980 of choice modelling, following the exclusively revealed-preference (RP) work described in the previous section, showed a strong trend towards stated preference methods, and in particular stated choice (SC) methods, to be used for leading-edge modelling work. There were many reasons for this, some of which we discuss below.

Another trend was the broadening of areas of application of choice modelling. While transport applications continued to be important, applications to other economic sectors became more common. For example, the well-known paper by Louviere and Woodworth (1983) used six examples to illustrate the points it was making on experimental design and analysis of SC responses: of these, three were transport examples, three from marketing.

These trends were very productive. SC methods opened many new opportunities for learning about the heterogeneity of preferences, as well as offering the opportunity to determine the values of non-market goods, to study new products and services, which cannot be observed in the current market, or to investigate valuations for attribute values outside the range currently observable. SC was also used simply to improve the efficiency of studies (e.g. on WTP) that had previously been based solely on RP data. The new work was able to build on the RUM foundations of theory, technique and software that had been laid in the previous years. Further developments of RUM methodology and software were made, in particular the introduction of the 'mixed logit' model that has become very widely used in recent years, aided by the increasing power of computers. It was not for several years, however, that non-RUM methods were actively explored by leading practical researchers, following the work of Tversky and Kahneman (1974), sometimes exploiting Prospect Theory (Kahneman and Tversky, 1979).

An area of theory that was of limited importance before the use of SC data was the need to deal with multiple responses from the same individual. Multiple responses help greatly in obtaining good information about heterogeneity, both between respondents and across the responses of one individual, while they also reduce error in the model estimation, improving the cost-effectiveness of surveys. A number of approaches have been developed to deal with multiple responses but issues remain. In many cases, analysts model inter- and intra-respondent variation explicitly, using the mixed logit approach (usually ascribed to Revelt and Train 1998, but see also Stiratelli et al., 1984). However, these models can

be difficult to interpret, because of the possibility that some dimensions of variation may 'proxy' for other dimensions that are not included in the model. For example, if the weight ascribed to one variable is allowed to vary in the population, it may take on the role of representing other preferences that vary in the population but are excluded from the model; including 'all' variations is of course impossible. Alternatively, a 'naïve' model may be used, without inter-personal variation, with the claim that the parameter estimates are consistent estimators of the parameters of the true model, but the proof of that claim is less well established in the literature than is generally thought (despite the huge number of citations of Liang and Zeger, 1986) and the procedures for calculating error measures for the parameter estimates are also subject to debate and can be onerous. Finally, some recent work attempts to construct models of choice for separate individuals, but again there remain a number of issues in this approach. The workshop discussion on these issues is reported in Daly et al. (Chapter 4, this volume).

In current choice modelling practice, the use of SC is dominant, to the extent that one may question whether revealed preference data is being unreasonably overlooked. The use of disaggregate RP responses in conjunction with SC information is reasonably well established, although some issues remain as discussed in section 2.2 below. However, there are several types of RP data that offer excellent possibilities for choice modelling, in conjunction with SC data or independently. Aggregate data types include the following.

- Simple counts of people (or vehicles, 'hits' etc.) entering or using shops, roads, entertainment facilities etc. are basic for validation or calibration of models, but also offer prospects for time-series modelling.
- Sales data can be collected in many contexts. In general this type of information offers little socio-economic detail but the large volumes and consequent reliability of the data give good prospects for modelling. The fact that expenditure and product type are available as well as simple numbers give much more information for modelling.

Disaggregate RP data can also be collected, for example in the following ways.

- Diaries of activities performed, trips made, appliances turned on, sites visited etc. can be used for quite detailed modelling, as in conjunction with collecting the diary itself, socio-economic data can also be collected.

- Brief 'intercept' interviews can collect information about what people are currently doing and a few pieces of background information. However, this type of information contains biases that are sometimes difficult to correct in modelling, because the sampling is usually related to both the choice made and exogenous data such as the respondent's location.
- GPS, like other automatic data collection, is still in its infancy as a basis for modelling. The large volumes of data that can be collected and the issues of privacy represent the most obvious difficulties, but there are other problems that also need to be overcome.

Data of these and similar types are often collected relating to transport behaviour, while analogous data exists for other sectors. The data is not always used for modelling, but can frequently be turned to that end.

One of the reasons that RP data was discarded in favour of SC experiments was the cost of collecting sufficient disaggregate RP data to match the (apparent) accuracy available from SC. However, much RP data – often hundreds of thousands of disaggregate records – is collected for other purposes and can often be obtained at minimal expense, subject to privacy guarantees. Of course, in this case the modeller cannot exercise influence over the design of the data, which therefore sometimes has important deficiencies, such as the omission of important data items, but often the trade-off is worthwhile.

The key advantage of RP data is that it records what people actually did. In SC data, there is always a doubt that the responses people give are representative of what they would actually choose to do in real circumstances corresponding to the hypothetical scenarios presented to them. Researchers have made significant progress in improving the reliability of SC responses, but the uncertainty can never be totally removed, simply because of the experimental nature of the interview.

The modeller must always be aware of the argument that, if RP data exists, then it is surely reasonable to expect the model to fit it. That is, unless the analyst takes account of the best RP data available, the model is likely to be obviously wrong. Moreover this failure is likely to be clear to everyone. Better, then, to accept the existence of the RP data and integrate it into the development of the model.

2.1 The nature of exogenous data for RP analysis

A major issue in the use of RP data is defining the variables and choice sets to which individuals are responding. In the case of SC, there is research on non-attendance to the attributes presented and on other effects, such

as 'halo', which might cause the respondent to indicate preferences different from his or her true preferences. In RP data, there has been very little research on such issues and the determination of 'correct' values for the attributes, or even what the attributes should be, becomes a matter of concern.

Early work using RP data, such as the 1960s transport studies in the UK reported in section 1.1, often asked respondents to report the attributes of their choice and the main alternative to it. Often under the highly misleading name of 'perceived' data, these attribute values were used in modelling. Unfortunately, what was not known at that time was that this type of reporting is subject to a series of biases and errors: rounding, self-justification, strategic responses of various kinds and simple ignorance, particularly relating to the unchosen alternative. Moreover, if it was required to use the model for forecasting, it would in principle have been necessary to create a model that would predict what respondents would report in the future context, when the attributes had perhaps changed. For this reason, RP analysis is more appropriately based on more objective calculations of the attributes.

For transport applications, objective calculation of the attributes usually implies setting up a network representing the transport system. Within the network, for each travel mode, routes can be determined for any origin-destination pair, most often using Wardrop's (1952) principle, and the travel times and costs calculated for these paths. This is not a simple process, as networks can be large and complicated, so that the quality of data becomes an issue, while the determination of the impact of congestion is also difficult, often requiring an iterative solution. However, with care, reasonably accurate measures of time and cost can be obtained. Crucially, these measures are reproducible, give the same result for travellers who did or did not actually choose specific journeys, and can also be obtained on the same basis in forecast scenarios when changes to network conditions may apply.

A more sophisticated approach is to accept that measurement error exists and to make appropriate provision for it. Walker et al. (2010) develop an application for transport modelling in developing countries, but error also exists, possibly to a lesser extent, in the data of economically developed countries, so that these procedures could well be extended to apply in other contexts.

For applications in other fields, depending on the nature of the data, analogous but hopefully simpler procedures can be imagined. Sometimes applications in other fields will contain a transport component, as the choice of location can be relevant for many kinds of goods and services.

The selection of attributes to include in the model depends on what is

relevant to behaviour and what can be measured cost-effectively. In many cases, it is not possible to measure all of the relevant characteristics and this may lead to bias when omitted characteristics are correlated with those included in the model. However, this is an aspect of modelling in general and applies to SC models as much as to RP: the omission of a variable from SC presentations does not mean that respondents necessarily ignore it. For example, if we are investigating the choice between trains and buses, SC respondents may take into account the fact that trains are generally more reliable and comfortable even if they are not given explicit information on reliability and comfort in the SC exercises.

The issue of choice sets is of concern in modelling with RP data, as there is rarely any indication of which alternatives were considered. Thus the model is generally one of choice from the feasible set. While it is generally concerning that we do not know which alternatives were considered before a choice was made, it is difficult to see how an explicit bias can be caused. Asking respondents which alternatives they considered may improve the accuracy of a model, but for forecasting we would need to develop a model of the formation of the consideration set, a challenge which does not yet seem to have been taken up to any great extent.

In summary, modelling with RP data necessarily implies a sort of 'reduced form' model, where objective measures proxy for both the true attributes and the true choice set. The advantage of RP data is that it reproduces real-life behaviour and that there is a forecasting procedure that can reproduce future-year 'data' compatible with the base-year data, so that forecasting for years well into the future is possible.

The further issue of adjusting the model to represent future behaviour also applies both to SC and RP data; further research is necessary on this point (see Fox and Hess, 2010). It appears that a key issue is to avoid overfitting, so that we can be more confident that variables appearing in the model truly represent real influences on behaviour.

2.2 Combining RP and SC data

An option that has been used in many studies is to make a joint analysis of SC and RP data, often using the error scaling approach devised by Ben-Akiva and Morikawa (1990), which was made operational for logit models by Bradley and Daly (1991). This approach facilitates the simultaneous use of different data types, hopefully correcting for specific weaknesses in each of the data types. For example, it is sometimes difficult to determine the true prices of the alternatives in RP data, while SC data may omit some attributes. Moreover, joint analysis makes it possible to include SC data in model systems being developed for forecasting (Daly and Rohr, 1998).

However, in some important practical studies it is necessary to sample respondents on the basis of the choices they have made, perhaps because this is the only reasonable contact procedure. There exist properly-based correction procedures to give consistent estimates for data sampled as a function of choice; there are also procedures, though complex, to deal with sampling which is partly choice-based and partly exogenous, which is a common feature of RP data. But when we sample SC respondents based on their actual (RP) choice, there is clearly a correlation between the RP choice and the SC responses, as the preferences relate to the same person. This case is of practical importance when we wish to predict switching to a new product – a salient example is predicting demand for a new high-speed rail service. Morikawa (1994) deals with the issue of correlation between RP and SP responses, an important part of this issue, but does not address the issue of sampling. It appears that there is no theoretical treatment of this problem in the literature.

The standard procedure for estimating WTP at present is to use SC data and this extends to the issue that is crucial for transport planning, the measurement of the value of time (VOT) in money units. In the most recent VOT estimations (Fosgerau and Hjorth, 2007; Börjesson et al., 2012) an important problem that arose, even when the analysis had been significantly improved over previous work, is that respondents gave clearly higher values per minute when asked to value larger time savings than when valuing smaller time savings. The average VOT obtained from the experiment, which is what is required for transport project appraisals, is then a function of the design, which is clearly unsatisfactory. There is also an issue concerning the measurement of time changes relative to current travel times (Daly et al., 2011). Here again the use of RP data appears to be an interesting approach to resolve some of the issues that have arisen.

An important context in which RP and SC data need to be used together is that of forecasting for new alternatives. For forecasting, the greater realism offered by RP data is essential, but when new alternatives or products are included in the forecast scenario SC may be the only reasonable approach. To apply SC results in a forecasting context it is necessary not only to take account of error scaling, but also there may need to be further adjustments to bring the SC model into the forecasting context (Daly and Rohr, 1998). A consideration here is whether the SC model predicts the share of each alternative *ab initio* or predicts switching from an imagined future in which the new alternative is not present. 'Inertia' effects, i.e. bias towards the currently chosen alternative, are also very relevant.

A further reason for relying on RP data in some circumstances is that some types of behaviour are not suitable for SC investigation. To make a successful SC interview, it must be possible for the respondent to imagine

the alternatives as being feasible. For example, in the transport sector, it would be difficult for respondents to imagine working in a different location (what sort of job would this be?) or taking a different number of holidays in the year (would I get time off?). These are issues that are much more realistically approached using RP data. In some cases, 'Stated Intentions' have been used, where respondents indicate what they might do, perhaps in rather general terms, in changed circumstances; while this approach has had some success, there have also been cases where the results have been less plausible and at present it seems that it should be used only when there are no good alternatives.

In summary, analysts need to be flexible when choosing the data type to be used for a given study. RP data has a number of advantages relative to SC and should be given more consideration than is currently the case in leading choice modelling work. In particular, combinations of SC and RP data can be very effective. Testing a model against RP is always advisable.

3. FORECASTING WITH CHOICE MODELS

A large part of the original motivation for choice modelling was to produce forecasts of demand for transport services. However, more recently the focus has been much more on obtaining estimates of WTP and on understanding the mechanisms influencing behaviour, in all the sectors of application of choice modelling. While these are important issues, the provision of accurate and defensible forecasts is also important for transport and many other important issues of public and commercial policy. It is, therefore, worth considering how choice modelling forecasting procedures work and whether these could be effective in wider contexts than those in which they are currently applied. There are interesting and challenging issues here.

Choice models predict the probability that an agent will choose each of the alternatives, conditional on the attributes of the agent and of the alternatives. In principle, all that is required to predict demand is to determine the number of agents of each type, the attributes of the alternatives they face and the amount they will demand of the alternative they choose. The key forecasting formula then gives the expected demand Q_j for alternative (product etc.) j by

$$Q_j = \sum_k w_k \cdot p_j(x_k) \cdot q_j(x_k) \tag{2.1}$$

where
w_k is the number of agents of type k;

$p_j(x_k)$ is the probability of choosing alternative j, given x_k;
$q_j(x_k)$ is the quantity demanded of alternative j, given x_k and that j is chosen;
x_k are the explanatory variables for agents of type k.

In a *discrete* choice model, the quantities demanded are unity, i.e. $q = 1$, and most choice modelling applications are of this form. However, the formula above, where amounts are not necessarily unity, is provided for completeness in dealing with models that mix discrete and continuous choice. Formula (2.1) appears to cover more-or-less all contexts in which choice models would be used for forecasting.

In the transport sector, in order to deal with the fact that transport infrastructure is expected to have a life of several decades, it is necessary to make long-term forecasts of behaviour. In other sectors, infrastructure may also be an issue, requiring forecasts for an extended period. In any case, the provision of services or marketing of products would be expected to operate over a considerable period and we need models that can deal with these circumstances. Forecasting over a period of any significant length means that we have to deal with changes in the population, i.e. changes in w, and to provide mechanisms for predicting how the attributes of the alternatives (x) will change.

3.1 Forecasting the population

Population forecasts were a necessary input to travel demand models from the earliest applications. As the models became more sophisticated, it became necessary to extend the forecasts, from simply the total population and employment in each traffic zone, to include zonal populations by segment: workers, car owners etc. Local planning agencies were often able to supply projections of the required information. However, as choice models began to be used more frequently for forecasting, the degree of segmentation needed increased further and transport analysts realised that they had to provide the required segmented forecasts themselves, perhaps based on aggregate numbers derived from other agencies, but with the details needed by the travel demand models derived by procedures devised by the analysts.

The population forecasting issues involved in travel demand, which are also relevant to forecasting in other sectors, are first to define the respondent types or segments, indexed by k in equation (2.1), and then to estimate w_k for each k. The definition of the segments obviously depends on the specification of the model, as it is necessary to define a separate segment whenever there are segment-specific variables in the model. Additionally,

it may be desirable to define additional segments to allow segment-specific processing of outputs from the model.

One approach that has been popular in advanced US travel demand models is to set $w = 1$ for all the segments and to define a segment for every member of the expected future population. However, while this simplifies the discussion and presentation of the model, it misses the opportunity to save calculation when population members are identical in respect of all the relevant attributes. It also introduces a specific problem, discussed below, when we need to expand a sample to represent a larger population. It may also imply a very large computational cost, which could be reduced if a sampling approach was adopted.

For travel demand forecasting, much use has been made of the technique of 'proto-typical sampling' (Gunn et al., 1983). The approach is based on the finding that, conditional on the marginal data provided by planning agencies, correlation between location and socio-economic variables samples is low. Samples are generated that, as far as possible, match the marginal aggregate data for each zone yet do not deviate too far from the original distribution of the population over key dimensions. The procedure used to achieve this balance is 'quadratic minimisation' or QUAD (Daly, 1998) which achieves a balance between the two objectives and provides a unique result with a limited amount of computer effort. The user can adjust the balance of the objectives depending on the forecasting context.

A different approach that has been used extensively in the US is Iterative Proportional Fitting (IPF). In this procedure, a sample of households is repeatedly factored to match marginal totals. It can be shown that this procedure also converges to a unique result without excessive computation. The difference between IPF and QUAD is that the former gives an exact fit to the marginal totals, so that inconsistency between these totals and the current distribution is decisively resolved in favour of the former.

It appears that the IPF approach has its relative advantage when the base-year data is good and the forecast does not differ too much from it. In contrast, QUAD deals reasonably well with a situation with more error. Further work is necessary to resolve these differences and to determine the optimum approach in different contexts.

There remains an issue with either approach if it is required that $w = 1$, as factoring methods such as IPF and QUAD naturally lead to fractional weights, so that it is not possible to get an exact match to the exogenous data with $w = 1$. Approaches to this issue can be computationally onerous and again further work is needed to determine what the best approach might be, perhaps depending on the circumstances of specific studies.

A particular issue is forecasting the distribution of income in the

population. Forecasts of overall income increase are fundamental to changes in many other variables, such as, in the transport context, car ownership. Income forecasts are often dubious and have political overtones, while deriving forecasts of income distribution presents a further difficulty. An approach is outlined by Daly and Fox (2012), while sensitivity testing is always advisable and further work would be desirable.

3.2 Forecasting exogenous inputs

The other side of the forecasting problem is obtaining forecast measures for x, the exogenous inputs. In principle this process is straightforward, but there are some potential issues.

An important issue may be that supply and demand are jointly determined in a market equilibration. Thus, if it is required to forecast the choices made by a population under changed circumstances, a solution must be found for this interdependence. In general, it is not reasonable to suppose that equilibrium will occur in these changed circumstances, or indeed that equilibrium exists in the current situation.

In travel demand modelling, an equilibrium solution is usually obtained for the demand model and the transport networks, which are mutually dependent because of the impact of congestion. This appears to be done primarily because it is not possible to define an alternative unique point, while a uniform, clear and precise definition of an evaluation point is necessary for the equitable comparison of different transport projects. The justification in terms of the demand model would then be that base year behaviour is modelled on the basis of a notional equilibrium network and future behaviour is forecast on the basis of a notional future equilibrium. The assumption is that the degree of disequilibrium in future will not be significantly different from the current situation and that errors introduced by disequilibrium in the base and future contexts will cancel out. This reasoning may sound tenuous, but devising a reasonable and better procedure is not easy, given the specific need for uniformity.

For other economic sectors where there is an issue of equilibration, the difficulties of making travel demand forecasts may serve as an example, perhaps an example of what should not be done. The detailed implementation of a procedure will depend on the particular circumstances of the forecasting context. In some cases, it may be necessary to construct a detailed model of the behaviour of suppliers as well as demand responses.

The calculation of elasticity values for a model is a useful validation that the model is behaving in a reasonable way in predicting changes in behaviour with respect to changes in exogenous variables. A simple calculation, using a forecasting system, is to change the value of an exogenous

variable, say the price of one alternative, and to observe the change in demand. Often, information will be available about elasticities in the relevant market from other sources. Note that this calculation must be made using an effective forecasting system, not just at the mean value of the x variables, as the latter will overestimate the elasticity. Specifically, for a logit model, it is easy to show that the expected mean sensitivity of demand, calculated over all the x values, is lower than the expected sensitivity at the mean by exactly the variance of predicted probability (Daly 2008):

$$\overline{\frac{\partial p}{\partial x}(x)} = \frac{\partial p}{\partial x}(\bar{x}) - var(p) \tag{2.2}$$

The elasticity will therefore be reduced proportionately. For other models, similar elasticity reductions will apply.

A final remark that may be made at this point is that the validity of the forecasting procedure depends on our belief that people will change their behaviour as a result of a change in the exogenous variables. Without this belief, there is no reason to suppose that the model is reflecting more than correlations observed in the base data, perhaps because demand is more fundamentally dependent on some variable correlated with a variable incorporated in the model. We must therefore base the model not only on its ability to reproduce observed behaviour, but its plausibility in explaining the behaviour in question. Ultimately, however, there is no guarantee that a model is reflecting correctly what would happen if exogenous variables change: we have some tests and some arguments that should indicate it is plausible, but no more than that.

3.3 General remarks on the forecasting function

The forecasting formula (2.1) can be interpreted as giving the expected demand for each alternative. In many cases the expected demand is exactly what is required – with perhaps some indication of the potential extent of error – for input to decision-making, whether formal or informal. However, an alternative approach, instead of using p_j, is to draw a random variable for each k and to use the value 1 with probability p_j and the value 0 with probability $(1 - p_j)$. It is clear that there is no bias between these approaches; the expected outcome, in a statistical sense, is the same. However, the approaches do have advantages and disadvantages in other respects.

Making a comparison between the expected-demand and sampling applications procedures for a specific model application is not always

easy, particularly as there are entrenched advocates of one or other approach. Moreover, there are many techniques for efficient implementation of each of the approaches which are known only to the relevant practitioners, so that comparison of the best implementations is difficult in practice. Initial independent comparisons are not yet complete (Algers et al., 2005). In particular, a clear comparison in terms of the computer time required for each approach has not yet been made convincingly. In any case this comparison depends on the number of samples drawn in the sampling approach and the degree of segmentation used in the expected-demand approach.

An important point in the comparison is the 'noise' present in the sampling approach: that is, if the procedure is repeated, with a different series of random numbers, a different result is obtained. The degree of difference also depends on the number of segments represented in the model, of course, as a greater number of samples drawn reduces the variance of most outputs. To set against this, the sampling approach offers a greater flexibility of output, since the records produced can be processed in many different ways. The importance of these characteristics will vary across specific study contexts.

The variation of the forecasts between different series of random samples cannot be taken as an indication of forecasting error or day-to-day variation, as might be thought. In the first place, this would ignore the correlation between days which is present in reality: many people go to work at the same place and by the same travel mode on successive days. Further, it would be necessary to assume that the model and its inputs are not the source of any important error, but we know that the model parameters are estimated with error, even if the model is correctly formulated, while the inputs to travel demand models (at least) are subject to considerable error (de Jong et al., 2007).

The introduction of the factor q in equation (2.1) extends the forecasting framework to include discrete-continuous models, first introduced some time ago (Train, 1986; de Jong, 1990) but more recently extended to cover choice of multiple alternatives (de Jong, 1997; Bhat, 2005). However, these papers focus on model estimation and the issues, if any, arising from the use of these models for forecasting have not yet been addressed.

4. FUTURE DEVELOPMENTS

I have argued that modern choice modelling would do well to revisit some of the issues that concerned early researchers in the field. Not that the many important advances made in recent years are in vain, simply that

some important achievements of early research deserve to be revisited and synergies with current work should be exploited.

For example, the earliest work on travel demand modelling used an entropy maximising approach which gave insight into model structures, yielding in particular the logsum formula. Now that it is understood that this approach is the dual of utility maximisation, it is worth asking whether there are other insights to be obtained from an entropy approach that could be applied to choice modelling in the RUM framework. This investigation would seem to fit with work giving an understanding of spatial choice (Cochrane, 1975; Daly, 1982; Jaïbi and ten Raa, 1998) and might help in understanding issues of spatial correlation and the competition of spatial attractions.

The main area of the early work I believe should be trawled for ideas relevant to modern work is the use of RP data. RP data is of particular relevance for forecasting applications, where the observation of real behaviour in the base situation gives some confidence that we may be able to predict real behaviour in the future. For WTP studies, RP data can also be helpful to overcome specific problems in SC experiments and simply to provide additional volumes of data, sometimes quite cheaply.

The focus of early researchers on issues of sampling is particularly relevant for the use of RP data (Manski and Lerman, 1975; Cosslett, 1981). After a long gap, further advances have been made recently (Bierlaire et al., 2008). These very sophisticated analyses give modern researchers the possibility of using RP data collected in a range of different ways. For example, data collected using in-car GPS used as a source for route or destination choice modelling would be biased towards higher income groups, who would be more likely to own systems, but might also be biased by the chosen route or destination, depending on the precise data collection method used. However, not all of the issues have been solved.

Similar but even more difficult issues arise with the analysis of SC responses gathered from respondents selected on the basis of their RP choices. The issue of choice being correlated to a limited extent with the sampling has not been addressed, as far as I know, while the issue is of considerable importance for the prediction of demand for new products such as high-speed rail services.

In forecasting, the issue of bias and error in exogenous data requires attention, as does the reconciliation of the various methods of population forecasting and identification of the cases where expected demand should be replaced by sampled forecasts.

An important task of choice modelling is to advise on values for non-

market goods and services. When we accept modern results on how people behave, e.g. that losses matter more than gains, or that functions are non-linear, we find severe problems in maintaining methods consistent with economic theory, which assume functions that are essentially linear. Here there is scope for some important research.

The hope in this chapter is that by giving a view of parts of the development of choice modelling and pointing out some of the key achievements of earlier researchers, we can understand better the problems we face and the tools available to address those problems. To summarise, despite the substantial advances that have been made in choice modelling, particularly since it was taken up on a larger scale beyond the transport sector, there remains a substantial amount of quite important basic research to be done. We have exciting prospects for the next few years to undertake this task.

REFERENCES

Alastair Dick and Associates (1978) Regional Highway Traffic Model, Department of Transport, London.

Algers, S., Daly, A.J., Kjellman, P., and Widlert, S. (1996) Stockholm Model System (SIMS): Application. In D. Hensher, J. King, and T. Oum, *Seventh World Conference on Transport Research*, Pergamon.

Algers, S., Eliasson, J. and Lars-Göran Mattsson, L.-G. (2005) Is it time to use activity-based urban transport models? A discussion of planning needs and modelling possibilities. *Ann. Reg. Sci.*, 39, 767–789.

Beesley, M.E. (1965) The value of time spent in travelling: some new evidence. *Economica*, 32, 174–185.

Ben-Akiva, M. and Lerman, S. (1979) Disaggregate travel and mobility choice models. In D. Hensher and P. Stopher (eds) *Behavioural Travel Modelling*, Croom Helm.

Ben-Akiva, M. and Morikawa, T. (1990) Estimation of travel demand models from multiple data sources. In M. Koshi (ed.) *Transportation and Traffic Theory*, Elsevier.

Bhat, C.R. (2005) A multiple discrete-continuous extreme value model: formulation and application to discretionary time-use decisions. *Transportation Research Part B*, 39 (8), 679–707.

Bierlaire, M., Bolduc, D., and McFadden, D. (2008) The estimation of Generalized Extreme Value models from choice-based samples. *Transportation Research Part B: Methodological* **42** (4), 381–394.

Block, H.D. and Marschak, J. (1960) Random orderings and stochastic theories of responses. In I. Olikn et al., *Contributions in Probability and Statistics*, Stanford University Press.

Börjesson, M., Fosgerau, M. and Algers, S. (2012) Catching the tail: empirical identification of the distribution of the value of travel time. *Transportation Research Part A: Policy and Practice*, 46, 378–391.

Bradley, M.A. and Daly, A.J. (1991) Estimation of logit choice models using mixed stated preference and revealed preference information. Presented to 6th International Conference on Travel Behaviour, Québec; published as pp. 209–231 in *Understanding Travel Behaviour in an Era of Change*, P. Stopher and M. Lee-Gosselin, Pergamon, 1997.

Cochrane, R. (1975) A possible economic basis for the gravity model. *Journal of Transport Economics and Policy*, January, 34–49.

Cosslett, S.R. (1981) Efficient estimation of discrete choice models, in C. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 51–111, MIT Press.

Daly, A.J. (1982) Estimating choice models containing attraction variables. *Transportation Research*, **16B**, 1.

Daly, A.J. (1998) Prototypical sample enumeration as a basis for forecasting with disaggregate models. In PTRC/AET Conference.

Daly, A. (2008) Elasticity, model scale and error. Presented to European Transport Conference, Noordwijkerhout, Netherlands.

Daly, A.J. and Rohr, C. (1998) Forecasting demand for new travel alternatives. In T. Gärling, T. Laitila and K. Westin (eds), *Theoretical Foundation for Travel Choice Modelling*, Pergamon.

Daly, A.J. and Zachary, S. (1975) Commuters' values of time. LGORU Report T55, Local Government OR Unit, Reading, UK.

Daly, A.J. and Zachary, S. (1978) Improved multiple choice models. In D.A. Hensher and M.Q. Dalvi (eds), *Determinants of Travel Choice*, Saxon House 1978; republished 2011 with Appendix.

Daly, A. and Fox, J. (2012) Forecasting mode and destination choice responses to income change. Presented at IATBR, Toronto.

Daly, A., Tsang, F. and Rohr, C. (2011) The value of small time savings for non-business travel. Presented to ETC.

Domencich, T. and McFadden, D. (1975) *Urban Travel Demand*, North-Holland.

Fosgerau, M. and Hjorth, K. (2007) *The Danish Value of Time Study*, available at http://www.dtu.dk/upload/institutter/dtu%20transport/pdf_dtf/rapporter/the%20danish%20value%20of%20time%20study_250208.pdf.

Fox, J. and S. Hess (2010) Review of evidence for temporal transferability of mode–destination models. *Transportation Research Record*, No. 2175.

Gunn, H., Fisher, P., Daly, A. and Pol, H. (1983) Synthetic samples as a basis for enumerating disaggregate models. Presented to P.T.R.C. Summer Annual Meeting.

Jaïbi, M.R. and Raa, T. ten (1998) An asymptotic foundation for logit models. *Regional Science and Urban Economics*, 28, 75–90.

Jong, G.C. de (1990) An indirect utility model of car ownership and private car use. *European Economic Review*, 34, 971–985.

Jong, G.C. de (1997) A micro-economic model of the joint decision on car ownership and car use. In P. Stopher and M. Lee-Gosselin (eds), *Understanding Travel Behaviour in an Era of Change*, Pergamon, Oxford.

Jong, G. de, Daly, A., Pieters, M., Miller, S., Plasmeijer R., and F. Hofman (2007) Uncertainty in traffic forecasts: literature review and new results for The Netherlands. *Transportation*, 34 (4), June, 375–395, Springer Netherlands.

Kahneman, D. and Tversky, A. (1979) Prospect theory: an analysis of decisions under risk. *Econometrica*, **47** (2), 263–291.

Lave, C.A. (1969) Modal split model. Thesis, Northwestern University.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
Lisco, T. (1967) The value of commuters' travel time. Ph.D. Thesis, Chicago University.
Louviere, J.J. and Woodworth, G. (1983) Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. *Journal of Marketing Research*, 20, 350–367.
Luce, R.D. (1959) *Individual Choice Behaviour*, Wiley.
Luce, R.D. (1977) The choice axiom after twenty years. *Journal of Mathematical Psychology*, **15**, 215–233.
Manski, C.F. (1975) Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, **3**, 205–228.
Manski, C.F. and Lerman, S.R. (1975) The estimation of choice probabilities from choice based samples. *Econometrica*, **45** (8), pp. 1977–1988.
McFadden, D. (1978a) The theory and practice of disaggregate demand forecasting for various modes of urban transportation. In *Emerging Transportation Planning Methods*, US Department of Transportation.
McFadden, D. (1978b) Modelling the choice of residential location. In *Spatial Interaction Theory and Residential Location* (A. Karlqvist, L. Lundqvist, F. Snickars and J. Weibull eds), North-Holland, Amsterdam.
McFadden, D. (2000) Disaggregate behavioural Travel demand's RUM side: a 30-year retrospective. International Association for Travel Behavior Conference, Gold Coast, Queensland, Australia.
McIntosh, P.T. and Quarmby, D.A. (1970) Generalised costs and the estimation of movement costs and benefits in transport planning. Mathematical Advisory Unit Note 179, Department of the Environment.
Miyagi, T. (1984) The conjugate dual approach to travel demand modelling. PTRC 12th Summer Annual Meeting.
Morikawa, T. (1994) Correcting state dependence and serial correlation in the RP/SP combined estimation method. *Transportation*, 21, 153–165.
Quarmby, D.A. (1967) Choice of travel mode for the journey to work, *Journal of Transport Economics and Policy*, 1, 273–314.
Revelt, D. and Train, K. (1998) Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of Economics and Statistics*, **80** (4), 647–657.
Stiratelli, R., Laird, N. and Ware, J.H. (1984) Random-effects models for serial observations with binary response. *Biometrics*, **40**, 961–971.
Thurstone, L.L. (1927) *The Measurement of Values*, University of Chicago Press.
Train, K. (1986) *Qualitative Choice Analysis*, MIT Press.
Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: heuristics and biases. *Science*, **185**, 1124–1131.
Vrtic, M., Fröhlich, P., Schüssler, N., Axhausen, K.W., Lohse, D., Schiller, C., and Teichert, H. (2007) Two-dimensionally constrained disaggregate trip generation, distribution and mode choice model: theory and application for a Swiss national model. *Transportation Research Part A*, **41**, 857–873.
Walker, J., Li, J., Srinavasan, S. and Bolduc, D. (2010) Travel demand models in the developing world: correcting for measurement errors. *Transportation Letters*, 2, 231–243.
Wardrop, J.G. (1952) Some theoretical aspects of road traffic research. *Proceedings*, Institute of Civil Engineers, Part II, Vol.1, pp. 325–378.

Warner, S.L. (1962) *Stochastic Choice of Mode in Urban Travel: A Study in Binary Choice*, Northwestern University Press.
Williams, H.C.W.L. (1977) On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning A*, **9** (3), 285–344.
Wilson, A.G. (1967) A statistical theory of spatial distribution models. *Transportation Research*, **1** (3).
Wilson, A.G. (1969) The use of entropy maximising models. *Journal of Transport Economics and Policy*, **3**, 108–126.