RESOURCE PAPER PREPARED FOR THE WORKSHOP ON:

**Methodological Developments in Activity-Travel Behavior Analysis**

**Modelling individual preferences,**
**State of the art, recent advances and future directions**

Elisabetta Cherchi

CRiMM – Dipartimento di Ingegneria del Territorio
Università di Cagliari (Italy)
Tel: 39-70 675 5274; E-mail: echerchi@unica.it
Introduction

“De gustibus non disputandum est” (Anonimous)

Despite the above famous statement, individuals have always disputed about individual tastes, and the decision making processes behind consumers’ choices has been a focal interest for decades. Although challenges against the theory of rational behaviour date back to the work of von Neumann and Morgenstern (1944), the dominating approach (at least in the transport field) has been the neoclassical economics assumption of rational decision makers (or even more extreme homo economicus), who always perform well planned and consistent activities, aiming to maximize some subjective measure of value (McFadden 1999).

The reason for this dominance is that economic theory has provided an elegant, rigorous and at the same time relatively easy to implement model, designed to describe individuals’ decisions and to provide quantitative forecasts with well-defined statistical properties. On the other hand, investigations in psychology, at least in the criticisms raised by the economists1, have had a propensity to generate lists of errors and biases and mostly fail (with excellent exceptions) in offering a coherent alternative to the rational-agent model (Kahneman, 2003). Psychologists recognise that this complaint is justified, at least partly, because intuitive thought cannot match the elegance and power of formal normative models. However, as Kahneman (2003) points out … “the alternative to simple and precise models is not chaos; psychology offers integrative concepts and mid-level generalizations, which gain credibility from their ability to explain ostensibly different phenomena in diverse domains”.

The origin of this divergence relies on the historically different views of the decision-making process that neoclassical economics and psychology have. While economists have been interested in mapping from information inputs to choice, treating the decision process as a black box, psychologists’ prime objective has been to understand what happens inside that black box: the nature of these decision elements, how they are established and modified by experience, and how they determine values. McFadden (1999) notes that what has made the distance between the two approaches even bigger is that psychologists view the decision process as dynamic and individual behaviour as local, adaptive, learned, dependent on context, mutable and influenced by complex interactions of perceptions, motives and attitudes. On the other hand, in the economic tradition preferences are primitive, consistent, and immutable (preference rationality), consumers behave as if they possess the formal tools with which to calculate the optimum adequately (perception-rationality), and the cognitive process is simply preference maximization, given market constraints (process-rationality).

The models that we (transport researchers) currently use to describe how people choose among a discrete set of alternatives are based on these assumptions of rationality in preference, perception and process. McFadden’s work (1978, 1981) on Generalised Extreme Value (GEV) formulation, which generalised the work of Williams (1977), provides a rigorous foundation for consumer choice modelling derived from economic theory. Although the original formulation of the random utility maximisation (RUM) as a behavioural model followed the economists’ theory of consumer behaviour, it also included “features of the taste template that were heterogeneous across individuals and unknown to the analyst, as well as

1 McFadden (1999) uses a powerful description for the feelings of those economists in front of the new evidence coming from psychology: “it has been like watching master carpenters to construct the scaffold for your hanging”.

Elisabetta Cherchi
Elisabetta Cherchi

unobserved aspects of experience and of information on the attributes of alternatives, interpreted as random factors” (McFadden, 2000). This led to the paradigm for generating discrete-choice models (DCM) commonly reported in textbooks (Ben-Akiva and Lerman, 1985; Ortúzar and Willumsen, 2001; Train, 2009), that the random part of the individual utility reflects the modellers’ lack of complete information about all the elements considered by the individual making a choice and the observed deviations of individual behaviour from perfect rationality (Tverski, 1972).

This paradigm posed the bases for the most important stream of research of the last 30 years. Since McFadden’s work, in fact, research activity in this field has been very proactive. Major progress has been made in exploring the potentiality of DCM to improve the ability to effectively reproduce individual behaviour. In particular, I would like to draw attention to two streams of research motivated by this work. The first refers to the microeconomic justification of DCM and, in particular, of the utility individuals associate to each discrete alternative. The second stream, and maybe the most productive one, has been concerned with the characterization of the error terms, and in particular the exploitation of the mixed multinomial logit (MMNL) model.

Research in both streams has aimed to improve the representation of the true phenomenon. However, the goal has always been that the measurable part of utility should be able to explain (as much as possible) the true behaviour in order to reduce the explanatory power left in the error term. This is correct except that, under the neo-classical theory, the systematic measurable utility was associated only with “rational” behaviour and what deviated from it was classified as error and hence minimised. Unfortunately, major improvements in model fit obtained with complex decompositions of the error term have given a strong signal that there are inherent limitations in the capability of microeconomic theory to explain individual choices and that we are still far from having a good representation (through known variables) of the real phenomenon. In fact, although RUM “takes a nod towards psychological theory” (Batley and Daly, 2006), the error term cannot be considered to properly explain behaviour that departs from perfect rationality. This is because errors are parameterized in terms of statistical distributions and the psychological concept of irrationality (i.e. not rational in the sense of neoclassical economic theory) is associated to the concept of randomness. As suggested by Ariely (2008), apparent irrationality can indeed be explained and predicted.

Illustrious scholars (McFadden, 2000; Ben-Akiva et al., 2002a) have strongly asserted the need to explore more seriously the suggestions provided by the psychological literature. After a shy start, the last decade has seen a surprising increase in the amount of work in this area (see for example Walker, 2001, Gärling and Axhausen, 2003; Bonsall, et al., 2007). Most of it has concentrated on demonstrating empirically that integrating psychology theory into the economic framework results in tangible improvement in terms of model fit, and interestingly most of it has been based on the MMNL structure. This is a key point, because the last years have witnessed also an increased awareness of the inherent limitations of the MMNL in terms of both estimation and especially prediction. In fact, notwithstanding the clear ability of this model to represent an ample range of behaviour via error term decomposition, several problems implicit in its structure have led analysts to lose confidence in the model. It is interesting then to understand whether or not these new models, which go beyond the rational postulate, still suffer from the above limitations or to what extent these are overcome.

In this paper I review research developments in the representation of the decision process, and for this I have divided it into two parts. The first is dedicated to reviewing the limitations of
the DCM and, in particular, of the MMNL model. Limits due to both the microeconomic theory of the rational user and to the exploitation of the error terms will be critically discussed. The second part of the paper reviews research belonging to the non-rational theory. Here I will concentrate mostly on those advances that still rely heavily on the DCM with the aim to discuss to what extent we are really moving forward with respect to the above limitations of the classical MMNL model. Although I will concentrate on research produced in the transport field, I will provide and rely on several references from the literature in psychology and behavioural economics.

I would like to mention that placing an accent on the limitations is not dictated by a pessimist view. On the contrary, it is intended as a proactive approach as it is precisely these limitations that form the starting point for and, above all, will stimulate new research. Another important consideration is that excellent reviews of both microeconomic theory (see McFadden, 2000; Bates, 2007; Jara-Díaz, 2007) and discrete choice models (see Ben-Akiva et al., 2002b; Ortúzar, 2006; Bhat, 2007) already exist, while a review of their limitations seems yet to be undertaken; at least this is what emerges from research conducted over the last few years.

I will conclude by discussing some open questions raised by the research conducted so far and giving some final thoughts about the amazing challenge unfolding before us over the next years.

1. Modelling human choice under the postulate of rationality

In the last 30 years analysts have spent much of their efforts in trying to improve the capability of GEV models to reproduce individual behaviour. The reason, as mentioned in the introduction, is that the main goal of transport engineers and economists, as opposed to psychologists, is to forecast demand and assess benefits. And, an important property of these models is precisely the successful marriage between an explicit theory of behaviour with a micro representation, allowing the constructive use of statistical goodness-of-fit measures for model specification and the derivation of consistent benefit measures (Williams, 1977).

The basic concept underlying microeconomic theory is the rationality postulate, according to which the consumer "always chooses the one he prefers from a set of available alternatives" (Varian, 1978). The concept of rationality used to define the homo economicus assumes that individuals possess a mental order of preferences that allow them to have perfect information about all the available options and the possible consequences of their actions. They are able then to associate to each option a utility function that measures the level of satisfaction they derive from it and to make finally their choice coherently with their preferences and with the constrains upon them. As reported in most textbooks in this area (I will use the notation of Ortúzar and Willumsen, 2001) given a certain set \( A = \{A_1, ..., A_j, ..., A_N\} \) of available alternatives, and a set \( X \) of measured attributes of the individuals and their alternatives, individuals:

1. Are endowed with a particular set of attributes \( x' \in X \) and in general will face a choice set \( A(q) \in A \);
2. Know all alternatives available in their choice set \( A(q) \in A \).
3. Evaluate each alternative \( A_j \in A(q) \) based on its characteristics \( x' \).
4. Associate to each alternative a level of satisfaction, measured through the utility index.
5. Compare the alternatives based on the level of satisfaction perceived and always choose the most attractive (i.e. the one that gives the highest satisfaction) subject to environmental constraints.

Formalising this theory, the choice of alternative \( A_j \) by the \( q^{th} \) individual will be observed if, and only if, the utility provided by this alternative is greater than the utility provided by any other alternative available to individual \( q \):

\[
U_{qj}(x'_{qj}) \geq U_{qj}(x'_{ij}) \quad \forall \ A_i \in A(q), \ i \neq j
\]  

(D1)

Differently from the Neo-Classical theory, as formalised in equation (1), the random utility maximisation approach assumes that the above “behaviourist” theory does not hold perfectly, or does not hold in a deterministic way in reality (Tverski, 1972). Moreover, even if the decision maker would know exactly her/his utility, the modeller is only able to observe certain components \( (x_{qj} \subseteq x'_{qj}) \) of the real vector of characteristics (Manski, 1977; Williams, 1977). Thus utility is treated as a random variable depending on the sub-vector of characteristics \( (x_{qj}) \) known and measurable by the modeller and at least one random component \( (\epsilon_{qj}) \) that captures everything that deviates from perfect rationality and/or all the relevant aspects of the phenomenon not explicitly known by the modeller:

\[
U_{qj}(x'_{qj}) = U_{qj}(x_{qj}, \epsilon_{qj})
\]  

(A2)

A key assumption for the derivation of the DCM is that random utility can be treated as the sum of the systematic, representative or observable part \( (V_{qj}) \), which is a function of known attributes \( x_{qj} \), and the random component:

\[
U_{qj}(x_{qj}, \epsilon_{qj}) = V_{qj}(x_{qj}) + \epsilon_{qj}
\]  

(A3)

Since for the modeller utility is a random variable, s/he does not know which alternative has the highest utility, but can only compute the probability that a given one will be chosen:

\[
P_{jq} = \text{Prob}\left\{ (U_{jq} \geq U_{iq}), \ \forall i \neq j, \ A_i \in A(q) \right\}
\]  

(B1)

However, from equation (3) we must assume that the residuals have a certain density function \( f(\epsilon) = f(\epsilon_1, ..., \epsilon_N) \), so the probability of choosing \( A_j \) can be computed as:

\[
P_{jq} = \int_{R_{N}} f(\epsilon) d\epsilon, \quad R_N = \begin{cases} 
\epsilon_{iq} \leq \epsilon_{jq} + (V_{jq} - V_{iq}), & \forall i \neq j, \ A_j \in A(q) \\
V_{jq} + \epsilon_{jq} \geq 0
\end{cases}
\]  

(B2)

where, depending on the distribution of \( \epsilon \), different discrete choice models can be generated.

In its original form, equation (3) assumed separability between the known and unknown parts of the utility, making the difference between what represents neo-classical rational behaviour \( (V_{qj}) \), and the random terms that capture the difference between true \( (U_{qj}) \) and measured utility \( (V_{qj}) \) very clear. This distinction motivated two major streams of research: the microeconomic justification of the representative utility and the characterization of the error terms.

Recent formulations, such as the famous MMNL model, assume that \( V_{qj} \) includes random elements also; hence it is not strictly correct to define \( V_{qj} \) as the known part of utility.
However, utility can still be derived from micro-economic theory with added errors. The MMNL utility function, in fact, is characterised by an error term with at least two components: one for obtaining the logit probability with the usual Extreme Value type 1 (EV1) distribution, and a second term that accounts for different components of unobserved heterogeneity, the distribution of which can be freely chosen by the modeller (Train, 2009). This second error term can be associated to known parameters, giving the random coefficients (RC) structure, or to unknown parameters, giving the pure error components (EC) structure. The most general utility for the MMNL model can be written as:

$$U_{qjt} = b_j x_{qjt} + \mu_{qjt} + b_j^\prime + \mu_{qjt}^{\prime} + \varepsilon_{qjt}$$  \hspace{1cm} (6)$$

where \( b \) are two vectors of taste parameters that might vary over alternatives \( A_j \) but are fixed over individuals \( q \) and choice situations \( t \), \( \mu \) are two vectors of unobserved components of the individual utility, \( x_{qjt} \) is as usual a vector of known attributes while \( z_{qjt} \) is a vector of attributes unknown (in value and nature) to the modeller; thus, without loss of generality they can be taken as equal to one for all alternatives or for groups of them. Finally \( \varepsilon_{qjt} \) is the EV1 error.

As well known, the great power of the MMNL stands in the freedom associated to the decomposition of both RC and EC random terms (\( \mu_s \)), which allows accounting for many patterns of substitution effects. In particular, specifying opportune the covariance matrix, the EC allows accounting for correlation among sub-groups of alternatives (Gopinath et al., 2005; Walker et al., 2007), cross-correlation, heteroscedasticity, dynamics and even auto-regressive errors (Hensher and Greene, 2003; Train, 2009). The RC version, instead, allows for the simplest random heterogeneity in tastes (Sillano and Ortúzar, 2005; Train, 2009) but also for correlation among different parameters in the same alternative and choice situation (Train and Sonnier, 2005), auto-correlation of the same parameter over choice situations (Gopinath et al, 2005) and heterogeneity across observations of the same individual (Hess and Rose, 2009). Greene et al. (2007) also include observed heterogeneity around the standard deviation of the random parameters and error component effects, specifying the standard deviation of the random parameter as a function of the individual socio-economic characteristics.

Other than the exploitation of the error term to accommodate complex patterns of taste heterogeneity and substitution, substantial developments have also taken place in the definition of the mixing distribution, in the analysis of identification issues, in the simulation methods to solve the multidimensional integral and in the algorithms to solve the maximum (for a review see Walker, 2001; Ben-Akiva et al., 2002a; Ortúzar, 2006; Bhat, 2007). However, and despite these major advances, some important problems still remain that make questioning the reliability of the DCM, and in particular of the MMNL, as appropriate tools to forecast demand and evaluate user benefits. Unfortunately, major improvements in the model specification (in both the systematic utility and the characterisation of the error term) do not guarantee that the model can effectively be estimated or that the estimated structure is reproducing correctly the true underlying phenomenon. Based on the recent literature, I will summarise the limitations of the MMNL model in the follows three points:

1. The difficulty to account for the correct microeconomic behaviour, i.e. the problem of the microeconomic derivation of the indirect utility function.
2. The difficulty to distinguish unequivocally the correct structure underpinning the available data, i.e. the presence of confounding effects.

3. The difficulty in discerning the true distribution; the characterisation of the error terms rely on the capability of the shape of the distribution to describe the true phenomenon.

I will discuss each point based on the findings in the recent literature and highlight the consequence in terms of estimation and prediction. Some questions for future research will then be raised and left open for the discussion.

Needless to say that all the problems mentioned above are compounded by the typical lack of large enough samples and of data of reasonably high quality. One such compounding is the so called empirical identification problem (Walker, 2002) whereby the model is estimable in principle but the information provided in the dataset is not sufficient to identify all the parameters. Chiou and Walker (2007) demonstrated that empirical identification problems are obscured if a low number of draws is used, while Cherchi and Ortúzar (2008) showed that empirical identification problems appear when a variable has low variability between alternatives.

1.1 Microeconomic derivation of the indirect utility function

Basically two approaches can be adopted for defining the utility function: the theoretical approach (Train and McFadden, 1978; Jara-Díaz and Farah, 1987), whereby the form of the utility function and the variables entering it is derived from microeconomic theory, and the empirical approach (Gaudry and Wills, 1978; Lerman and Louviere, 1978) whereby the most appropriate functional form is just that producing the best fit to the data, provided minimal theoretical conditions are not violated.

The theoretical approach soon became more attractive than the empirical one since it is consistent with the mathematical structure of the random utility models that are derived precisely from assumptions about consumer behaviour. However, the neo-classical theory is quite strict and the way it is made operational for the practical purpose of model estimation is at the root of many problems in estimation (e.g. confounding effects) and in the description of the decision process (evidences of non-rationality).

The classical microeconomic formulation proposed by Train and McFadden (1978) states that individuals derive utility only from the total amount of goods \(G\) they can purchase and the time they can spend in leisure activities \(L\), subject to income and time constraints:

\[
\max U(G, L) \quad \text{s.t.} \quad G + c_j \leq E + \omega W \\
L + W + t_j = T \quad \forall j \in M
\]

where \(E\) is unearned income, \(\omega W\) represents the income people can earn if they work \(W\) hours at a wage rate \(\omega\), \(c_j\) and \(t_j\) represent respectively travel cost and travel time by mode \(j\); \(T\) is the total time available, excluding the minimum time required to sleep and other life-compulsory activities, and \(M\) is the number of discrete alternatives. The maximum utility individuals can attain subject to the above constraints represents the indirect utility function conditional on mode \(j\):

\[
V_j = U\left((E - c_j), (T - t_j), \omega\right)
\]
Equation (7), or the analogous formulation proposed by Jara-Díaz and Farah (1987) for fixed income\(^2\), refers to homogenous individuals and this is a major limitation that inevitably produces inconsistency with the model estimated in practice. In fact, strictly speaking, different problems should be formulated for different socio-economic groups\(^3\), and consequently different discrete choice models (i.e. at least different systematic utilities) should be estimated for each sub-group. However, due to practical problems, samples are never homogeneous. This is why socio-economic (SE) attributes are always included in the utility function, though without a consistent microeconomic derivation. Train and McFadden (1978) assumed that the SE variables act as proxy for attributes unobserved in the cost and time constraints. However, in this case we are not truly capturing the systematic heterogeneity. Other authors (Algers et al., 1996; Dillen and Algers, 1999) assumed that SE characteristics are elements of the direct utility function though it is not clear how people derive utility directly from SE characteristics, and even more how they would maximise them.

Theoretically, SE attributes should influence the direct preferences (\(\beta\)) for goods and leisure (e.g. \(U = kG^{\beta_1}L^{\beta_2}\) in the typical Cobb-Douglas function), which is consistent with the classical assumption of systematic heterogeneity in preferences for attributes, although it imposes constraints on the variability of this heterogeneity.

No matter the specification of the microeconomic formulation, other important problems arise when the indirect utility function (8) is made operational to match random utility theory. Here, commonly a Taylor expansion is firstly made to transform any general indirect utility function into a linear in the parameters (albeit not in the attributes) utility function, then only the part of the utility relevant for choice is considered; then, once the truncated utility function is defined, a randomly distributed error term is usually added to account for unobserved variables, measurement errors and uncertainty.

Using a Taylor expansion is not a problem in itself, but truncating the expansion to only the first or the second order terms has several important consequences. Firstly, at least a second order expansion is required to account for the effects of the time and income constraints in discrete choice (Jara-Díaz and Videla, 1989). It is interesting to note that these constraints actually set a limit to the possibility of trading-off attributes, which in a certain way allows the model being not perfectly compensatory. However, only Hultkrantz and Mortazavi (2001) use this formulation explicitly to account for non-compensatory effects. This effect is lost when a linear approximation is used, because income and time constraints cancel out in the comparison between pairs of alternatives.

Another important consequence is that all terms of the Taylor expansion left out because of the truncation engross the error term, and this might be at the root of some confounding effects as previously discussed. Cherchi and Ortúzar (2001), for example, found that omitting second-order variables from the utility function add commonality to the alternatives that appeared erroneously perfectly independent of one another. Amador et al. (2008) showed that omitting the cost squared term might lead to erroneously detecting random heterogeneity in preferences.

---

\(^2\) More complete (and complex) formulations have been proposed that assume other elements influencing direct preferences (Troung and Hensher, 1985; Bates, 1991; Gunn, 1996; Kraan, 1997; Jara-Díaz, 1998; Jara-Díaz and Guevara, 2003 Karlström et al. 2007). But they still refer to homogenous individuals.

\(^3\) Train and McFadden (1978) formulation is, in fact, directed to freelance workers, while that of Jara-Díaz and Farah (1987) to employees. Gunn (1996) proposed different formulations for workers, no-workers and housekeepers.
Finally, as noted by Batley (2008), the translation of RUM to practice reveals a problem of compatibility in the analogy between the Neo-Classical and Random Utility models; in contrast to its theoretical underpinnings, the specification of discrete choice RUM in practice yields models that exhibit the properties of cardinal utility. In line with this argument, Amador and Cherchi (2009) demonstrate how taking this class of order-preserving transformations can lead to misinterpretation of the econometric results, such as detecting randomly distributed and correlated parameters and/or income and time effects which are in fact not present.

As discussed in the introduction, the microeconomic approach is now under severe criticism, because several analysts have produced evidence about the inherent non-rationality of individual behaviour. However, not everybody agrees that the microeconomic theory is completely wrong and/or that we are currently in the position to abandon it. Despite the major criticism against DCM they are still the most powerful tools available to make predictions. Some questions then arise:

− Should we abandon this theory? Or, are there circumstances where individuals still behave rationally, at least partially?
− Which of the current problems faced when estimating DCM can be attributed to inherent non-rational behaviour rather than an approximation in making the microeconomic theory operational?
− Should (and/or could) we think of different microeconomic formulations that relax some of the assumptions of perfect rationality? This question is related to the new developments that will be discussed in Section 2, as they mostly include elements of non perfect rationality to the indirect utility function.

1.2 Confounding effects

By confounding effects we refer to the difficulty of unequivocally identifying the “correct” effects underpinning the observed process. Confounding effects means that although different structures can produce good and equally likely estimation results they cannot be estimated jointly because the data are not sufficient to identify all the parameters (Walker, 2001; Cherchi and Ortúzar, 2009).

The difficulty of the DCM to distinguish between different model structures is an old problem raised almost 30 years ago by Williams and Ortúzar (1982) and Daly (1982). But, it became more severe with the advent of the MMNL model, because the possibility to decompose the error terms comes at the cost that there can be many different (identified) factor analytic specifications that provide approximately the same ‘best’ fit to the data (Walker, 2002).

The literature reports quite a long list of cases of confounding effects, mostly occurring among random terms. Hensher (1998) found that specific alternatives may produce spurious correlation if unobserved heterogeneity is ignored. Swait and Bernardino (2000) showed that taste heterogeneity may be potentially confounded with heteroskedasticity and preference trees. Munizaga and Alvarez-Daziano (2001, 2005) noted that reproducing a Nested Logit (NL) model with the MMNL is not so simple due to its inherent heteroskedasticity, while Amador and Cherchi (2009) showed that the common practice of disregarding the scale factor of the overall utilities (based on the order-preserving transformation) might affect the
parameter estimates producing confounding effects, such as detecting erroneously randomly distributed and correlated parameters. Hensher and Greene (2003) suggested that heterogeneity is a special type of correlation, amongst choice situations, while Hess et al. (2005) found that modelling jointly correlation in the choice set and random taste heterogeneity can be problematic, and that the problem grew in severity with the complexity of the model (as they moved from a mixed NL model to a mixed cross-correlated logit). Bhat and Castelar (2002) referred to the likely interactions among inter-alternative error structures; scale differences between revealed and stated preference data; unobserved heterogeneity and state dependence. In this line Smith (2005) commented that one should be cautious in interpreting random parameters if researchers are unable to model state dependence. Nevertheless, he also stated that if a more elaborate parameterization of preference heterogeneity is used, excluding state dependence may magnify the apparent preference heterogeneity in the model but not necessarily generate it where it does not exist. This could be viewed as the converse of the problem explored by Heckman (1981), where the emphasis was on the emergence of spurious state dependence if heterogeneity was not modelled properly (Ortúzar, 2006).

As discussed recently by Cherchi and Ortúzar (2009), most of the above confounding effects are indeed due to the compensatory nature of the random terms in the MMNL. They show that the property that only the difference between alternatives matters in conjunction with the compensatory rule, may lead to another confounding effect: between correlation and heterogeneity in tastes and response, which can appear in estimated models and produce misleading forecasts. Using real and simulated data, they found that a significant specific random parameter may reveal correlation among competing alternatives rather than variation in tastes, cautioning that this is important if the model will be used as a forecasting tool.

In terms of forecasting errors, an even more serious problem may occur when confounding effects appear between error components and some important micro-economic effect. In this case the problem is that we can end up supporting the wrong policy. For example Amador et al. (2008) found that randomness in the cost parameter was actually masking the presence of income effect. The effect can also appear in the opposite direction. For example, Amador and Cherchi (2009) refer to the possibility that not accounting for heterogeneity might erroneously “reveal” the presence of income and time effects.

All these effects pose serious question regarding the confidence we can put into our modelling results, maybe not as much in estimation but certainly in prediction and as a guide to choose appropriate policies. Cherchi and Ortúzar (2009) report that there seem to be no statistical tests that can unequivocally help in spotting the right model specification and recommend the pragmatic solution of estimating alternative specifications and compare their results, looking carefully at the absolute value of the random parameters, and the relative values of the alternative specific constant and correlation coefficients, apart from all the statistical tests available. Some authors (Horowitz, 1980; de Jong and Daly, 2006; Hess and Daly, 2009) have suggested to compute confidence levels for the measures derived from the estimated models; in the case of predictions, this will soon become a requisite in work for private agencies such as construction companies and banks involved in public-private partnerships, where

---

4 They use the Likelihood Ratio test proposed by McFadden and Train (2000) to determine if mixing is needed, as well as the marginal effects and marginal square error (MSE) tests.

Elisabetta Cherchi
Accountability is a real issue. However, all these methods are currently prone to bias and more exact measures are difficult to come by\textsuperscript{5}.

A number of comments, and related open questions, can be drawn from this literature review:

- Most confounding effects are found among random structures. Should we continue to put so much effort into decomposing the error terms? How much confidence can we place in the capability of reproducing true behaviour if the model relies so heavily on the error terms?
- Do some methods exist that allow us to identify the true phenomenon (is it just a matter of research on that)?
- Would better data (more information and of better quality) solve the problem, or at least help to do so? Or should we simply move away from the intrinsic structure of the DCM?
- Are new studies and models that account for non perfect rationality, as we will discuss in Section 2, free of confounding effects?

1.3 Choice of the distribution

Choosing the correct distribution to reproduce the heterogeneity underlying population preferences has been one of the major research interests in the past years, but it is still a central problem. Hensher and Greene (2003) comment that “selecting the distributions for individual parameters is challenging enough”. Ortúzar (2006) reports that “an extensive literature exists on this subject, but the question of choosing the correct distribution seems still open”. In a paper appeared last year, Train (2008) comments that unfortunately there is no guidance “for finding the mixing distribution that attains an arbitrarily close approximation. In practice, researchers have tended to specify a parametric distribution and estimate its parameters testing alternative distributions. However, whatever distribution is used, dissatisfaction with the properties of the distribution soon surface”.

The problems associated with the most usual distributions are well known, so I will only summarise the issue. For example, as the Normal distribution allows parameters to be either positive or negative, it is difficult to understand if the proportion of the population reproduced with a wrong sign is by construction (i.e. because of the shape of the distribution), because they are wrongly coded or (worse) answered untruthfully (see the discussion in Sillano and Ortúzar, 2005) or even because individuals really have opposite tastes (see the discussion and related papers in Hess et al., 2005; Cirillo and Axhausen, 2006). In real cases almost all attributes have associated a parameter which is logically bounded, i.e., it can be only positive or negative or, at least, it cannot be unboundedly large. So, Lognormals which avoid wrong signs have been proposed but also heavily discredited because they are usually harder to estimate, mask the problems of errors in the data and, having a long upper tail, have high probabilities of yielding large portions of cumulative mass close to zero (Bhat 1998; 2000; Revelt and Train, 2000; Sillano and Ortúzar, 2005; Meijer and Rouwendal, 2006).

Bounds can be applied to Normal distributions (Train and Sonnier, 2005) but this may be problematic. Firstly there are no rules on how to set the bounds and secondly this is likely to produce skewed distributions, which have the problems of overestimating the true mean toward the bounds, distorting the welfare measures (Cherchi and Polak, 2005), and are more likely to give empirical identification problems (Cherchi and Ortúzar, 2008). Triangular

\textsuperscript{5} This comment appears in the forthcoming fourth edition of Transport Modelling by Ortúzar and Willumsen.
distributions avoid unboundedly large values but do not prevent wrong results for those parameters that have a defined sign. Hensher and Greene (2003) proposed to constraint the distribution, making the spread or standard deviation of each random parameter a function of the mean. However, there appears to be no strong theoretical basis for imposing such constraints, so they conclude that, except for the sign of the parameters, there are no theoretical arguments to support one distribution over another. Moreover, such a specification uses one parameter for two purposes which can be overly restrictive (Train, 2008).

Train and Sonnier (2005) proposed a more flexible form of bounded distribution, the Johnson SB distribution, which has the advantage that its density can be shaped like a Lognormal with an upper bound and with thinner tails below the bound, but it is more flexible as can be shaped like a plateau with a fairly flat area between drop-offs on each side, and can even be bi-modal. Moreover, the bounds of the SB distribution can be estimated as parameters, rather than specified by the modeller, though at a cost that identification becomes an issue since the difference between the upper and lower bounds is closely related to the variance of the latent Normal term. Notwithstanding its nice properties, the SB distribution has rarely been used. The reasons are not clear but few works have been published where this distribution has been tested (Hess et al., 2005 and 2006; Fosgerau and Hess, 2009) and they report that use of SB distributions led to slow convergence, might have overall fit lower than the Normal distribution and can give problems with parameter significance. Finally, Fosgerau (2006) compared sixteen parametric distributions with nonparametric techniques. He found that although the SB distribution was the most adequate choice for the purpose of estimating the mean parameters, many distributions turned out to be suitable choices for his data since the predicted probabilities using these agreed closely with the observed probabilities over the observed range of the index.

Fosgerau and Bierliare (2007) note that using only the goodness-of-fit to compare models does not allow one to reach clear conclusions about the validity of the random parameters distribution and propose a test based on semi-nonparametric techniques to decide if a given distribution is appropriate or not. However at present the test is able to discriminate only between Normal and Lognormal distributions and only when one parameter is random.

To overcome the problem of the choice of the distribution, one of the most recent and promising advances adopt non-parametric models where a family of distributions is used to approximate the variation of tastes in the population (Dong and Koppelman, 2003; Burge and Rohr, 2004; Fosgerau, 2006; Train, 2008; Fosgerau and Hess, 2009; Bastin et al., 2009). However, the flexibility of nonparametric methods comes at the cost of requiring more parameters than in most parametric specifications. Nonparametric methods require large datasets and the accuracy of the approximation rises with the number of parameters and that increases the possibility of empirical singularity at some iteration (Train, 2008). Moreover, while the nonparametric specifications are generally superior (in terms of goodness of fit) to their parametric counterparts, this superiority is not as pronounced when the true distribution has large mass points at zero. And more importantly, from a benefits estimation perspective nonparametric specifications might perform worse than the parametric ones (Cherchi et al., 2009).

Finally, it is important to mention that several papers have affirmed the importance of estimating the parameters at the individual level rather than to be satisfied with the population distributions (Huber and Train, 2001; Train, 2001; Brownstone, 2001; Sillano and Ortúzar,
2005; Godoy and Ortúzar, 2008; Greene et al., 2005). However, individual parameters are seldom computed. As before, the reasons are not evident.

In the light of the above literature review, some questions arise:

- As non-parametric distributions require even more data and draws than the parametric ones, are they really “feasible” for work of practical dimensions?
- Are there cases where parametric distributions can still be used? Or should we abandon them? If so, what is the validity of the many papers that still adopt parametric distributions? Note that the new developments discussed in Section 2, rely mostly on the Normal distribution.
- Can non-parametric distributions be easily used in more complex cases (such as when there are many random parameters and full correlated covariance matrices), where even the parametric distributions show their limits?
- Why, despite their clear benefits, are individual parameters seldom computed? Why, after a moment of fame, does the subject seem to have been abandoned?

2. Beyond the rationality assumptions

Over the past 30 years, research in behavioural decision making has provided a large body of evidence that people very often exhibit departures from rationality. Research in cognitive psychology, marketing, sociology and more recently in behavioural economics has systematically (but also extensively and unequivocally) demonstrated that individuals very often violate all the assumptions of the rational postulate. In particular:

- Individuals do not have well-defined (i.e. primitive, consistent, immutable, regular) preferences. Contextual details (even objectively irrelevant), attitudes and personality, habit and inertia, existing anchors, loss aversion, social interactions, other-regarding, all affect individual behaviour in systematic ways.
- Individuals are not “perfect machines” able to process any amount of information to maximise overall utility. Individuals do not always invest the required degree of effort in evaluating alternatives and have limitations in their capacity to process information and make calculations. People are adaptive decision makers who possess a repertoire of different strategies in making judgments and choosing among alternatives.

Behaviour deviating from perfect rationality has been often defined as “anomalous”, with a sort of negative flavour. As Gowdy (2008) comments … “it is ironic that a large body of evidence suggests that “lower” animals act more in accordance with the economic model of rational choice than humans do. (...) Far from describing higher-order, complex human behaviour, the axiomatic rational choice model strips away everything that makes humans unique as highly intelligent social animals”. This is why the rational choice model is inappropriate to describe all but the simplest kinds of human decision making (Gintis, 2007).

---

6 Almost all human behaviour has a substantial rational component, at least in the broad sense of rationality (McFadden, 1999).

7 Behavioural Economics is the name given to the new discipline at the crossroads between economics and psychology (or more properly a branch of psychology called behavioural decision research). Behavioural Economics has its roots in the work of von Neumann and Morgenstern (1944) and aims to integrate psychological insights into formal economic models.
Compared to other fields, it seems that transport researchers have been more devoted to the rational economic theory. However, over the last couple of years research interest has significantly shifted in favour of psychology theory and an increasing amount of evidence of departure from rationality has been produced, including choices in transport. A major review of the vast literature, mainly but not only, in transport, has allowed me to identify the following among the most important themes considered:

1. Asymmetry between willingness to pay and to accept
2. Attitude and personality
3. Habit, inertia
4. Non compensatory choice and other heuristics
5. Learning and dynamic processes
6. Choice under risk (prospect theory)
7. Compromising effects (regret theory)

In this paper I will focus only on the first three issues, because I prefer to limit the analysis to a few points to have the possibility of trying to discuss them in more detail. Furthermore, I chose those issues where there was a more extensive use of the DCM described in Section 1 and, of course, where also the problems discussed would presumably apply. In this sense, both regret theory (see Chorus et al., 2009) and prospect theory (Bonsall et al., 2007; Avineri, 2004; Avineri and Bovy, 2008; Gao et al., 2009) are based on different paradigms from RUM. My choice also reveals a preference for the mode choice context rather than route choice or activity-scheduling/participation where learning and dynamic processes related to the formation of inertia have been addressed for quite a while (Mahmassani and Chang, 1985, 1986; Mahmassani and Herman, 1990; Mahmassani, 1997; Avineri and Prashker, 2006; Jotisankasa and Polak, 2008). Finally, regarding non compensatory behaviour and other choice heuristics, several interesting papers have produced evidence in the modal choice context (see for example, Cantillo and Ortúzar, 2005; Castro et al. 2009; de Lapparent, 2009; Swait, 2009). However, theoretical work comes mostly from the marketing literature and there are already some recent reviews and interesting analyses of the problem (see for example Adamowicz et al., 2008 and Denstadli et al., 2009).

Although each issue above has often been studied to account for a specific behavioural anomaly (according to the classification by McFadden, 1999), it is not always possible to attribute the result to only one specific effect; as the decision-process is a highly complex system where components are mutually interrelated. I will try to highlight all the behavioural effects behind each specific effect.

Finally, although I will make several references to work produced in the general literature of behavioural studies, I will not review them, as this area is vast and far beyond the scope of this paper. Excellent reviews can be found in Camerer (1999), McFadden (1999), Rabin (1998; 2002) and Maital (2007). A list of key readings in behavioural economics was compiled by Camerer et al. (2003) along with a discussion of policy implications of bounded rationality.

2.1 Asymmetry between WTP and WTA

Although it is difficult and maybe not even useful, to trace back the first works in transport that have departed from the rational theory, surely among the first examples are the bunch of studies on the asymmetry between willingness to pay (WTP) and willingness to accept
These are two user benefit measures typically derived from microeconomic neoclassical theory. In particular, the \textit{WTP} is the measure most used in transport and it is derived precisely from the DCM discussed in the previous section. Understanding the reasons why these two measures differ is then crucial.

Both measures became very popular at the beginning of the 90s. A broad review of about 30 years of studies on differences between \textit{WTP} and \textit{WTA} can be found in Horowitz and McConnell (2002). At that time it was not common to talk about bounded rationality. Reasons for these differences were then searched mainly in the microeconomic theory or attributed to problems on practical calculus, due to the well-known limits implicit in the hypothetical experiments used to elicit this type of information. For example, in the late 90s (Bradley and Gunn, 1990) concluded that the large ratio between \textit{WTA} and \textit{WTP} reflected short-term behaviour, because people were not able to re-arrange schedules and thus substantial increases in travel time could cause disruption. McFadden (1997) pointed out that when stated preference data were used, the overestimation of the \textit{WTA} could be due to the consumer mistrust of hypothetical trade offers or could arise from strategic misrepresentation, since the consumer may see an opportunity to gain substantial rents by overstating \textit{WTA}.

From the economic point of view, since the concepts of \textit{WTP} and \textit{WTA} are the equivalent unitary measure of the compensating (\textit{CV}) and equivalent variation (\textit{EV})\footnote{The \textit{WTP} measures how much people would give up, when getting the benefit or avoiding the loss, in order to feel just as well off as they did before. The \textit{WTA}, instead, corresponds to the minimum amount the individual would be willing to accept to forgo the reduction in some characteristics (or goods).}, the presence of income effect\footnote{The common use of \textit{WTP} is as equivalent to \textit{CV} and conversely the \textit{WTA} as equivalent to \textit{EV}. However, this equivalence only applies when the policy changes produce an improvement (less price or better quality). If the policy changes imply worse conditions, then \textit{WTP} corresponds to \textit{EV} and \textit{WTA} to \textit{CV}.} is the first natural reason behind the differences between \textit{WTP} and \textit{WTA} values. However, Hanemann (1991) pointed out that if the public good has almost no substitutes there is no reason why \textit{WTP} and \textit{WTA} could not differ vastly. Hence their difference depends on income effect but also on a substitution effect. In line with this, Horowitz and McConnell (2002) found that high \textit{WTA}/\textit{WTP} ratios depended on the type of goods and in particular were highest for non-market goods, next highest for ordinary private goods and lowest for experiments involving forms of money (such as lotteries). Income effect and, to some extent, also substitution effects are the reasons of the differences between \textit{WTP} and \textit{WTA} found by Borisova and Goodman (2003), in a study on the value of travel time to get to a health centre; although they also recognised that uncertainty about individuals’ preferences could contribute to the higher \textit{WTA} amount.

But differences in \textit{WTA} and \textit{WTP} can also be explained on the basis of psychological aspects such as the asymmetries between individual attitudes to gains and to losses. Studying individual choices under risk, Kahneman and Tversky (1979; 1984) concluded that apparently the pain of marginal losses (loss aversion) exceeded the benefit of comparable gains; hence, individual choices were more responsive to anticipated losses than to equal and opposite anticipated gains. However, the gap between \textit{WTA} for less of a commodity and \textit{WTP} for more of this commodity might depend also on the consumer’s initial position. This \textit{reference dependence} suggests that decision makers care about changes in outcomes as well as about the final outcomes themselves. Hence people are more sensitive to changes that are coded as
losses (relative to a reference point) than to equal-sized changes that are perceived as gains (Ho et al. 2006). A classic example that demonstrates reference dependence and loss aversion is what Thaler (1980) called *endowment effect*. Experiments by Thaler and Johnson (1990) and Kahneman *et al.* (1990) established that endowment effects were not only pervasive and substantial, but also almost instantaneous, so that they do not come from sentimental attachment to long-term possessions.

The question of to what extent loss aversion in money, in addition to loss aversion in specific goods, contributes to observed disparities between the *WTP* and *WTA* valuation of goods, has controversial answers. Bateman *et al.* (2005) conducted an adversarial collaboration experiment between Kahneman *et al.* (1990), who predicted that loss aversion would not occur in buying tasks, and Bateman *et al.* (1997), who predicted that it would. The evidence from their experiment favours the latter prediction, although not decisively. In fact, they concluded that loss aversion in goods, loss aversion in money, and tactical and cautious heuristics may all be implicated in the familiar disparity between *WTA* and *WTP*. And the problem is that it is very difficult to eliminate these effects in experiments and, in particular, in contingent valuation studies.

Hultkrantz and Mortazavi (2001) recognised that the *WTP/WTA* disparity is due to a combination of income and substitution effect with psychological effects such as preference uncertainty, reference points and cognitive thresholds. Interestingly, they used a second order Taylor expansion of the indirect utility function to account for the possibility of a *WTP/WTA* disparity. They found that both *WTP* and *WTA* were bounded, which they say reflect a rather simple underlying set of choice rules employed by the average respondents, though they were not able to distinguish this effect from others, such as the effects of the real social cost.

Lately and also thanks to the popularity of behavioural studies, research on *WTP/WTA* has rediscovered a new interest, focusing on loss aversion and reference point effects, although experimental effects seem still to play an important role. For example, Bateman *et al.* (2009) show that differences between *WTP* and *WTA* can be generated by imperfect comprehension of, or uncertainty regarding, a good. They argue that this increases response variability, which in turn lead survey respondents to resort to heuristics (such as loss aversion) to formulate responses. Such strategies would, in turn, raise measures of *WTA* compensation for losses relative to *WTP* for gains. Lanz *et al.* (2009) suggests that an extremely high ratio *WTA/WTP* (even 1000) could be attributed mainly to loss-aversion in all attributes. In particular, they justified the pronounced degree of gain-loss asymmetry in the price attribute, as further evidence that monetary outlays are perceived as losses rather than foregone gains, in line with the reference dependent preferences. Hjorth and Fosgerau (2009) show that loss aversion may depend on socio-economic characteristics (such as age, education, income, and gender) and on experimental design. They also found a significant negative effect of congestion on loss aversion in both the time and cost dimensions\(^\text{11}\), which they explain as due to learning effects. People who travel in congested circumstances are more used to variable travel times, although it could also be an effect of the experimental design. Knetsch and Wong (2009) provided further tests and found that reference states matter, and that the consequences of differing procedural control manipulations are likely to be, in large part, due to their impacts on the reference states perceived by participants. Further, these findings appear consistent with the role that reference states play in the choices people make in their everyday lives. Finally

\(^{11}\) They use data from a highly abstract stated preference route choice experiment with only two variables (time and cost). It would be interesting to see if the same effects appear also in a more realistic setting.

Elisabetta Cherchi
Hjorth and Fosgerau (2009) argue that loss aversion would not be a problem if we were convinced that it is a rational phenomenon reflecting true preferences; by which they mean the preferences corresponding to the well-being that people experience as a consequence of their choices. Fosgerau and De Borger (2009) provide a theoretical demonstration that loss aversion may be given a rational basis, in the sense that it leads to maximal expected hedonic utility. They show that if the degree of asymmetry is known, the hedonic marginal rate of substitution is identified from choices exhibiting loss aversion providing an explicit link from hedonic to choice preferences.

In a somehow counter tendency, some papers report that recent natural and laboratory experiments have shown that observed valuation disparities might be due to well-known incentives of standard theory rather than to any asymmetry in preferences over gains and losses (Knetsch and Wong, 2009). These findings seem in line with previous work done by Plott and Zeiler (2005; 2007) who found that when procedures to control for incentives that classical preference theory predicts to be relevant are used, this brings about a dramatic elimination of the large disparities between gain and loss valuations (that were evident when these controls were absent). They suggest that factors such as the experience of individuals in making similar transactions and the incentives induced by experimental designs may well have an influence on people’s apparent valuations of gains and losses in particular instances.

Notwithstanding major advances in this field, it seems that some important questions remain open:

- It is still not clear whether WTP and WTA are really different or if the methods used to measure them are prone to errors.
- If they differ, it is still not clear to what extent this difference depends on classical economic effects or on psychological effects. This question relates to the more general question of to the extent individuals behave according to the rational postulate or not?
- On the other hand, the WTP and WTA are based on the concept of trade-offs between cost and some other attributes. Hence they implicitly assume compensatory behaviour, according to the rational postulate. Can we still consider them valid measures?
- It seems that none of the studies examined accounts for random heterogeneity in tastes. Could the difference between WTP and WTA be partly explained by random heterogeneity?

### 2.2 Attitudes and personality

Attitudes have been defined as stable psychological tendencies to evaluate particular entities (outcomes or activities) with favour or disfavour (McFadden, 1999). Attitudes and personality are major factors in determining motivation and the structuring of cognitive task. They also influence perceptions and they may receive feedback from process and choice, as the decision-maker reconciles and rationalizes trial choices. Attitudes are formed over time and are affected by experience and external factors that include socioeconomic characteristics. They can be any latent characteristic of a decision-maker and thus incorporate concepts such as memory, awareness, tastes, goals and capabilities etc. (Walker, 2001).

The effect of attitudes and personality in transport choice has been a concern for long time but recently it has received another major boost thanks to recent developments in hybrid discrete choice modelling with latent variables. In the transport field, researchers have used various techniques in an effort to explicitly capture psychological factors in choice models. However,
most applications have used Structural Equation Models (SEM) or latent variables incorporated into discrete choice models.

SEM have been mainly used to specify and test alternative causal hypotheses between attitudes, perceptions, stated behavioural intentions and actual behaviour. A review of this work is reported by Golob (2003). Papers include attitude towards carpooling (Golob et al., 1997), environmentally friendly policies (Golob and Hensher, 1998; Sakano and Benjamin, 2000), financial support for public transport (Levine et al., 1999), fairness and infringement on personal freedom (Jakobsson et al., 2000) and residential location (Choocharukul et al., 2008; Scheiner and Holz-Rau, 2007).

Choo and Mokhtarian (2004) used factor analysis to derive variables representing attitude, personality and lifestyle, while Domarchi et al. (2008) included attitudes, habit and affective appraisals as explanatory variables through dummy variables. As reported by Walker (2001), these fitted variables contain measurement error, and to obtain consistent estimates the choice probability must be integrated over the distribution of the latent variables, where the distribution of the factors is obtained from the factor analysis model.

Nowadays, the recommended approach to incorporate attitude and perceptions into discrete choice is the latent variable models where the latent factors and the discrete choices are jointly estimated, either sequentially or simultaneously. The inclusion of subjective elements into discrete choice models has been a topic of discussion since the beginning of the 80s (Ortúzar and Hutt, 1984; McFadden, 1986). Nevertheless, latent variables models only became popular among transport researchers much later with the work of Walker (2001), who presented a general framework and methodology for incorporating latent variables into choice models via the integration of choice and latent variable models and the use of psychometric data.

Walker (2001) discussed two applications in a transport context. One was based on the work of Morikawa et al. (1996) who estimated a mode choice model between car and rail that incorporated two latent variables, ride comfort and convenience (this was extended later by Morikawa and Sasaki, 1998). The second example refers to the work of Polydoropoulou (1997), who estimated a more complex model involving more latent variables and combining a RP dataset with two SP datasets. In both examples, the results showed that latent attributes had significant parameter estimates and that their inclusion resulted in a large improvement in the goodness-of-fit of the DCM. Further examples of almost a decade ago are the seminal piece of Ben-Akiva et al. (2002b) and the paper by Morikawa et al. (2002).

A renewed interest in latent variable models spread across transport researchers in the last few years. Bolduc et al. (2005) represent the first example of the analysis and implementation of a situation characterized by a large number of latent variables and a large number of choices. They discuss the estimation problems associated with the large dimensional integrals that arise in their case and suggest simulation driven Bayesian and classical approaches to the econometric estimation of these models. Vredin Johansson et al. (2006) studied the effect of five latent variables (comfort, security, flexibility, convenience and environmental awareness) on a typical modal choice (among bus, car and train). Differently from previous studies, they used individual specific, not mode specific, latent variables to explain choice, which means that they did not construct latent variables for non-chosen modes to avoid endogeneity problems. In line with all previous findings, this application confirmed that latent variables were indeed crucial to explain mode choice, although, in their experiment, only environmental preferences, comfort and flexibility were found significant. Dannewald et al.
(2007) suggest that values are important to determine the classic attitudes towards mode choice, such as preferences for comfort/convenience, flexibility and safety. They proposed a hierarchical hybrid choice model where respondent values (power, hedonism, and security) determine the individual choice criteria (flexibility, possession, passivity, and environment protection) which in turn influence their trip mode choice. Vásquez Lavín and Hanemann (2008) studied how psychometric information can help to characterize taste distribution among individuals with a MMNL model. They proposed a method that integrates psychometric measures and revealed preferences to describe heterogeneity in preferences for congestion in a recreation demand models. In the particular case of congestion, the traditional MMNL model (with both Normal and Lognormal distributions) gave results that were inconsistent with the way people feel about the site attribute.

Some applications have referred to residential location. For example, Kitrinou et al. (2009) studied four latent variables (perceptions about the quality of life in island areas, satisfaction of the commuting trip, satisfaction of the current residential area and geographical mobility) that recognize the impact of policy variables referring to accessibility of the island area (with regards to both the transport and the telecommunications systems of the area). While Walker and Li (2006) use latent class models to study lifestyle groups and how lifestyle affects the choice of residential location.

Some of the most recent work in the field has dealt with estimation methods. While in the past, the sequential (Morikawa et al., 1996; Vredin Johansson et al., 2006) and the simultaneous (Polydoropoulou, 1997; Morikawa and Sasaki, 1998; Dannewald et al., 2007; Bolduc et al., 2008; Bolduc and Alvarez-Daziano, 2009) methods have been used without too much questioning, comparison between the two methods has recently become an important issue in the transport literature. Ben-Akiva et al. (2002b) argued that the sequential approach should result in more efficient estimators of the involved parameters, while Bolduc et al. (2008) showed that although the simultaneous estimation of hybrid models requires the evaluation of complex multidimensional integrals, simulated maximum likelihood offered an unbiased, consistent and smooth estimator of the true probabilities. Raveau et al. (2010) make a critical analysis of the results obtained from the empirical application of the sequential and simultaneous methods to real\textsuperscript{12} and synthetic data. They found that both methods were unbiased and resulted on estimators that are not statistically different; however, fairly different point estimates were obtained when computing the value of travel time. Although these differences are less serious than results omitting the latent variables, they still could have severe consequences when forecasting or evaluating transport policies based on the results of the alternative models. Moreover, they pointed out that any subsequent problems that may arise from the use of a sequentially estimated model, instead of a simultaneously estimated one, can not be attributed to information omission (which is the case of the model without latent variables), but to information misuse when obtaining the estimators. So, they suggested that the simultaneous approach should be preferred, although the method is more complex and lack of available commercial software might render difficult its diffusion as estimation method. Finally, Bolduc and Alvarez (2009) demonstrate that a Bayesian simultaneous estimation offers an enormous improvement over the classical estimation, because the choice model is only a part of the whole behavioural process in which we now incorporate individual attitudes, opinions and perceptions, thus yielding a more realistic

\textsuperscript{12} The real data refer to a modal choice context for urban trips to work with many choice alternatives (the majority of previous studies dealt with just two options).
econometric model. They apply the model to a modal choice context including two latent variables (environmental concern and appreciation of new car features) and found that this improved representation outperforms standard discrete choice models because allows one to build a profile of consumers and their ability to adapt to technological innovation with regard to sustainable private vehicle alternatives. However, they comment that further research is needed to generalize the method proposed.

Walker (2001) commented that “the early signs indicate that the methodology is promising”. All the evidence provided in these years confirm that using latent variables provide indeed crucial information to policy-makers and transportation planners applying new transport systems and in particular developing sustainable transportation systems. These results in fact suggest that although the typical level-of-services attributes (time, cost, frequency and so on) are of course relevant, the significance of attitudes and personality in mode choice model, confirm that there are other ways to attract individuals to the, from society’s perspective, desirable public modes of transport.

The general conclusion is that future mode choice models can be more powerful if they take into account individuals’ attitudes and personality traits. However, an important problem is that these latent variables are not directly susceptible to policy intervention, because attitudes and personality traits cannot be easily forecast. Walker (2001) warned that “for specific applications it would also be useful to conduct validation tests, including tests of forecasting ability, consequences of misspecifications (for example, excluding latent variables that should be present), and performance comparisons with models of simpler formulations”.

Yañez et al. (2009c) are one of only two examples of an analysis of latent variables models in forecasting. They found that the market share changes predicted by hybrid models were larger, but not systematically, than those of classical MMNL models for all alternatives. This may indicate that the hybrid models are effectively more sensitive, as we would expect, but this higher sensitivity did not imply just a simple amplification of the effects involved. They concluded that this helps to reveal the importance of including latent variables in choice models. The latent constructs can not be used normally for prediction purposes, as they are not based on objective factors; however, as the latent variables are not observed it is still necessary to integrate the choice probabilities over their whole variation range, conditioning them by their explanatory variables (i.e. subjective attributes such as comfort and reliability). As level-of-service variables do not enter the latent structures they are not able to capture changes in these types of variables. Yañez et al. (2009c) warn that this is an important requirement to take into account if the aim is to forecast using this type of models. Finally, Raveau et al. (2010) also discuss this problem and suggest examples, but conclude that these issues still need to be explored in more depth.

Notwithstanding the above major advances, a couple of questions arise:

- Following a comment reported in Walker (2001), could some of these major improvements in fit be captured in the choice model by including in the base choice model the additional variables that are included in latent variable structural model?
- As most cases discussed above are based on the typical linear in parameters formulation, to what extent the problem discussed in the previous section apply also to latent variable models?
- One of the major benefits of latent variable models is the improvement in model fit. However, learning from the experience gained with the MMNL (see section 1), what can
we say about the capability of latent variable models to correctly reproduce the true underlying phenomenon?
- It is true that before validating a model it is important to understand how to estimate it; but it would be desirable not to make the same “mistake” made with the MMNL, where in spite of the plethora of literature on estimation sufficient evidence in terms of validation and forecasting is still lacking.

2.3 Habit and inertia

In opposition to the economics tradition that assumed the existence of preferences, psychologist have always argued that preferences are constructed on the spot whenever needed, based on the task and context factors present during choice or preference elicitation (Payne et al., 1993; Slovic et al., 1990). However, a recent trend among behavioural researchers is to postulate that consumers have inherent preferences (a favoured combination of attributes) and through trial and error learn what they like; preferences then can change and stabilize over time. Basically people “acquire a taste” over time. However, what translates a first decision into long-term habit is the mechanism that led us to repeat that first decision without thinking again about the reasons (“ideal trade-offs”) why we behave in such way.

The influence of habit (leading to inertia) in the choice process has been largely discussed in the literature. The first studies are dated back to the end of the 70s (Goodwin, 1977; Blase, 1979; Daganho and Sheffi, 1979; Williams and Ortúzar, 1982). Since then, researches have always shown interest for this issue, though not always with the same intensity. To a certain extent, attributes such as car availability, miles travelled, seasonal public tickets, number of trips per week (typically used in cross-sectional mode choice models) are indicators of habit (Gärling and Axhausen, 2003), though of course do not explicitly relate the current choice with the previous ones. Mackie et al. (2001) found that an inertia term introduced into the utility function was able to account for different evaluation of losses and gains in a SP exercise. An “inertia dummy” has frequently been considered to control for the effect of the current choice in mixed RP/SP estimation (Morikawa, 1994, Bradley and Daly, 1997; Cherchi and Ortúzar, 2002). Recently Cantillo et al. (2007) proposed a model where inertia between RP and SP data is a function of the previous valuation of the alternatives.

Other interesting works have studied inertia using panel data, where data are gathered at different point in time around changes to the external environment (i.e. the supply or the socio-economic characteristics). Golob (1990) employed three waves of data (one year apart) from the Dutch National Mobility Panel to analyse travel behaviour stability. Using structural equation models he found inertial and lagged relationships between income, car ownership, car mobility and public transport mobility. Hirobata and Kawakami (1990) included inertia to predict traveller mode switching due to changes in transport levels-of-service in the short-run. Inertia was defined as a function of the level of service of the attributes before the change. Bradley (1997), using before and after data, estimated dynamic logit models that accounted for response lags and state dependence in order to study the effect on mode choice of a new rail commuter line. Swait et al. (2004) proposed a model that includes prior behaviour on a time-series context and compared it with typical static model. Importantly they pointed out that policy changes or environmental changes have temporal impacts. Thus, certain changes (e.g. a loss in environmental quality) will remain in the memory of the consumer for several periods and will affect choice and welfare for several periods, even after the effect has been eliminated. Chatterjee and Ma (2006) used a four-wave panel to examine the time-scale of
behavioural responses in changes in travel modes and monitored the change in attitudes and behaviour. They found that four months after the change the rate of new users was low but stable, but then it declined sharply. However, they also observed that in the same period the awareness of the service, positive attitude towards it and overall usage has grown slowly. Thøgersen (2006) used structural equations modelling to study to what extent the current behaviour toward public transport was influenced by past behaviour, current attitudes and perceived behavioural control. His analysis confirmed that past behaviour is the best predictor of future behaviour; however the panel study showed that attitudes towards using public transport and perceptions about its ability to fulfil one’s transport needs were influenced positively by the use of public transport, and that the more people used public transport the more likely it was that they would like it up to the point that they changed habit.

Srinivasan and Bhargavi (2007) used a dynamic mode choice model to account for rapid and substantial changes in the fast growing Indian economy. Their results indicate that accounting for dynamics yield better model fit and more credible estimates of potential improvement measures. Moreover, improving the level-of-service alone would not produce the anticipated benefits to public transport agencies, as it failed to overcome the persistent inertia captured in the state-dependence factors. Yañez and Ortúzar (2009) and Yañez et al. (2009a) studied the effect of inertia and shock on individual behaviour as a function of the valuation of the alternatives in previous waves of a panel; they used a four-wave panel where the first was gathered just before and the remaining three 1, 6 and 18 months after the implementation of the radically new and initially much maligned new public transport system in Chile. They found that the inertia effects vary among individuals and over options, but not over waves. Moreover, they highlight that the inertia effect might represent a disposition to change (increasing utility value) or, in case of unattractive options, might reinforce the habit (i.e. increasing the probability of maintaining the usual choice).

Some evidence about inertia was found also using short -or continuous- panel data (i.e. data gathered over a continuous period of time). For example Ramadurai and Srinivasan (2006) estimated a MMNL mode choice model that accounts for habit persistence and state-dependence between two consecutive days. They found that accounting for these effects not only produced a substantial improvement in model fit but these models also outperformed others in validation tests. Simma and Axhausen (2003), using a multi-day panel over a 6-weeks continuous period, found that travel commitments (car ownership and public transport season tickets) in one period affected mode usage in the next period. With the same dataset, Cherchi and Cirillo (2008) estimated a modal choice model where inertia was indirectly measured through the day-to-day repetitiveness of the same choice. They found that at least 50% of the improvement in the overall model statistics were due to the presence of repeated observations. However, Yañez et al. (2009b) warn that having identical observations increases the efficiency of the estimated parameters in the MMNL only because this increases the sample dimensions, but it reduces the capability to reproduce true phenomena.

Although (most of) these works explicitly account for the effect of previous choices (i.e. choices made in previous periods) in the current choice, they all assume that individuals still make a trade-off (i.e. go through a decision process) for any choice in any period. To what extent individuals actually repeat the decision process every time (i.e. every wave) is a matter of discussion. According to Hoeffler and Ariely (1999), consumers clearly have some form of preferences but they do not go through the extensive effort of constructing their preferences for every decision. They learn from past decisions and adapt their consumption behaviour.
However, preferences are more likely to be constructed when encountering a new domain, which justifies the specification of the trade-offs to explain part (being the inertia the other part) of the choice after a change in the environment (such as in most of the long-term panel data). The question, “how big should be the change to induce individuals to perform in a decision process” does not seem to have an answer yet (witness the shock effect in Yañez et al., 2009a). Although this is an important issue, if individuals do not construct their preference, then the trade-off should not be included in the waves after the first one. Interestingly Garvill et al. (2003), in a field experiment, showed that forcing individuals to think about their choice released the effect of habit, confirming that the effect of habit is that individuals do not choose the “best” option.

This effect is actually what psychology calls self-herding. Ariely (2008) argues that habit-process formation starts with a first positive impression of something that we tried for curiosity or necessity, rather than because we engaged in the “ideal trade-off”. Since preferences are reinforced by positive experience, if we like it then we tend to repeat the same action even if it is not the “ideal” choice. Moreover, individuals believe that something is good (or bad) on the basis of their own previous behaviour (self-herding), hence the more they repeat an action, the more they feel that they are acting on the basis of their preferences, up to a certain point when they have already made this decision many times in the past, so that they assume that this is the way they want to behave (their preference).

On the other hand, if there are not external changes the effect of habit is that individuals repeat exactly the same choices, which is precisely what has been found in continuous panel data. Several other studies have examined the repetition (or analogously the day-to-day variability) of the individual travel-activity pattern finding quite different results. Some studies report that the frequency of daily trips was remarkable stable from week to week (Pas and Koppelman, 1986; Kitamura and van der Hoorn, 1987; Cherchi and Cirillo, 2008); some others have found, instead, that identical complete daily patterns are seldom repeated (Hanson and Huff, 1986; Huff and Hanson, 1990; Pas and Sundar, 1995). Others have stayed in a middle ground, concluding that behaviour is neither totally repetitious not totally variable (Schilich and Axhausen, 2003). Whatever is the proportion of repeated observations, in these cases the “rational” choice (i.e. the evaluation of the attributes) should be used only to represent the first time an individual makes the decision (the anchor), while subsequent choices should not be explained anymore by this “ideal” process. However, this does not mean that repeated choices should be disregarded. As demonstrated by Hoeffler and Ariely (1999), the act of making a choice and repeated choices both led to increased preference stability; then repeated choices should play a role, though it is not clear how they should be treated. This problem has been approached by Yañez et al. (2009b), who tested different ways of including repeated choices in model estimation. Though their results seem to depend on some statistical properties of the model, the issue of how to account for this behavioural effect is certainly important.

Another interesting issue is whether inertia influences the overall preference for utility (as commonly assumed) or if it affects the preference for specific attributes. Experiments conducted by Hoeffler and Ariely (1999) show that individuals develop stable preferences by learning the importance of different attributes and using them consistently. Individuals learn the weight to assign to each attributes, simply experiencing trade-offs among attributes (Fischer and Hawkins, 1993; Hawkins, 1994; Tversky et al., 1988).
While the above models can be used to predict changes in individual behaviour, several researches report about field experiments were the effect of habit is studied on selected groups of individuals after the implementation of some specific measures. Bamberg et al. (2003), Fujii and Kitamura (2003) and Thøgersen and Møller (2008), report that a free travel card effectively releases the grip of habit on car drivers’ travel mode choices, and makes them carry out intentions about using public transport to a significantly higher extent than at the baseline. However, different results were found regarding the long effect of the interventions. Contrary to Fujii and Kitamura’s result, Thøgersen and Møller found that the use of public transport fell back to the level of the control group after the promotion period ended, when participants had to pay a full fare. They comment that this result shows that an economic incentive in the form of a free travel card is not necessary for breaking the influence of habit on travel mode choice. Meloni et al. (2009) examined the short-term changes in activity-travel behaviour resulting from a cap-and-trade-like program (where individuals have fixed credits to spend in kms driving a car and those virtuous, who fall below their cap, can sell credits to those exceeding their limits). Their initial results suggest that, opportunely motivated, individuals reduced (though slightly) their trips by car in favour of walking while increasing chained trips.

Bamberg et al. (2003) highlight the importance of a good public transport system for breaking the habit. It is interesting to note that in these experiments, the type of incentive may also play a role. In fact, offering something for free has a “special” effect on people which is not consistent with the amount of the discount offered. For example, Ariely (2008) shows that reducing a price from \( 3X \) to \( X \) is less effective than reducing the price from \( X \) to zero. The reason seems to reside into the human fright for loss. There is not visible possibility of loss when we choose a free item. Finally, Seethaler and Rose (2006) argue that information-based campaigns and incentives are insufficient and sometimes counter indicative to breaking habit, and suggest using persuasion techniques that include reciprocation, consistency, social proof, linking, authority and scarcity.

Summarising, studies on inertia have produced a good variety of approaches and interesting results, but the impression is that much work is still needed. In particular, a couple of questions that seem important are:

- Contrary to much evidence, almost all our models assume that inertia explains only part of the current choices (i.e. individuals still make trade-offs of the current attributes). How can we determine to what extent individuals make a trade-off in the current choice? Can we really answer this question with our models or might there be confounding effects that prevent us from disentangling this effect?
- Since habit starts from the first time individuals make their first choice, should we trace back the first time individuals make a choice based on the trade-offs? How long should our panel be to get this information? Does knowing how people make the first choice help? Or should we concentrate on how people change behaviour now?
- Are panel data currently available suitable for really capturing inertia and more importantly for predict how to break it?
Conclusions

The prime objectives of transport analysts should probably be to estimate and evaluate the effects produced by policy interventions. For this reasons we have been mainly interested in tools that allow us to forecast demand and assess benefits. However, demand is formed on the basis of individual choices, and benefits depend on how well off individuals are after an intervention. Understanding how individuals choose among different alternatives and what drives their preferences is then a crucial issue. In the last decades, the parallel work of economists and psychologists has depicted two different pictures of the individual decision process: the rational choice of the economists based on perfect information and evaluations of the outcomes, and the bounded rational choice of the psychologists based on limited knowledge and the incapacity of evaluating all the elements that matter in the choice. Economic theory became dominant not only among transport researchers because it was able to provide an elegant, rigorous and at the same time relatively easy to implement modelling framework, designed to describe individuals’ decisions and to provide quantitative forecasts with well-defined statistical properties. While psychologists mainly concentrated in providing empirical evidence against the overreaching use of economic theory.

Today it seems that psychologists have succeeded as using psychology insights to understand individual choice processes represents almost the prevalent field of research. In recent years the amount of work in this area has increased in a surprising way. Interestingly, researches in transport have provided evidence for almost all the anomalies associated with the rational behaviour paradigm. In my review, I have focused on some of them (asymmetry between willingness to pay and to accept, and the effect of attitudes, personality and habit), but evidence has been produced also on the effect of the choice context, the effect of social interactions and social norm. Also evidence exists that preferences are made sequentially over time and are not always self-regarding, that individuals do not use only one single decision rules, that they do not always maximise overall utility, do not always choose what seems to be in their best interest, do not always invest the required degree of effort in evaluating alternatives, and that they may have limitations in their capacity to process information and make calculations.

Not all of these “anomalies” have received the same degree of attentions. Some (such as those I reviewed in the paper) have been studied in depth; for some others, one or very few papers exist. And certainly, one task for the future is to provide enough evidence for all the above anomalies. A more challenging task, though, is to try instead to study some anomalies jointly in an attempt to disentangle their relative effects. Until now, most analysts have focused on only one particular aspect of non rationality at a time. However, individual decision processes are highly complex and involve all or many of the above issues together. The risk involved in studying one effect at the time is that confounding (such as that experienced in classical discrete choice modelling) might occur also among anomalies.

Another important point, is that most of the work in transport (and in particular that which I focused on in this review) has concentrated on demonstrating empirically that integrating psychology theory into the economic framework results in tangible improvement in terms of model fit. This is in line with the general tendency, in many areas, that supports the integration between psychology and economic theory. This aims to take advantage of a well defined economic theory, improving it with the integration of psychological elements, and...
Elisabetta Cherchi recognises that both components of rationality (perfect and bounded) can play a role in understanding decision processes.

Unfortunately, the label “well defined economic theory” has been the object of some questioning. For example, there is an increased awareness of the inherent limitations of discrete choice models in general, and of the mixed logit model in particular, in terms of estimation and especially prediction. In spite of the clear ability of the model to represent an ample range of behaviour via error term decomposition, several problems implicit in its structure have made analysts lose some confidence in it. In particular, major difficulties in reproducing true behaviour have been demonstrated. These problems could be attributed to microeconomic theory (i.e. following the psychologist’s suggestions, individual behaviour deviates from perfect rationality). However, an extensive literature in this field shows that problems also depend on the way the theory is made operational, to match random utility theory, in relation to the assumptions required to characterize the error term (i.e. relying on random distributions) and on the intrinsic structure of discrete choice models (i.e. equally likely structures are possible and this generates confounding effects). These problems, of course, might also occur if new models that go beyond the rational postulate are used.

Moreover, all these problems are compounded by the typical lack of large enough samples and of data of reasonably high quality. Unfortunately, the requirements of good data increases with the complexity of our models. The new formulations integrating psychology and economic theory extend the behavioural capabilities of the model, but are even more demanding in terms of good quality data.

Given the above limitations of discrete choice models, some have started to suggest that they should be abandoned. Although, this is not likely to happen, it is certainly a sign that more work is needed to improve the confidence we can pose in these models. Improving the quality of our datasets and the knowledge about the phenomenon under study is obviously and as ever, a good starting point. Potentially, many model structures are prone to confounding effects and relying only on the model results to reveal the presence of many complex effects might be risky. At the same time, we still need more research to define formal methods that help identifying the true underlying phenomenon. This is true for models under the classical rational theory, but it seems as important for the new breed of models derived to account for bounded rationality.

On the other hand, it seems important to keep firmly in mind the achievements already obtained. The literature review on the limitations of discrete choice models, suggests that advances once considered important tend to be forgotten when we embark on a slightly different direction of research (I am referring, for example, to the results on the distribution of random heterogeneity, to the need to carry on validation other than estimation, to the importance of testing the model capability to reproduce the true phenomenon). If this is true for models under the classical rational theory, it seems as important for the new breed of models.

Finally, if both components of rationality play a role in the decision process, it seems important to analyse under what circumstances and/or to what extent individuals evaluate alternatives under the perfect rationality rather than the bounded rationality approaches. This is crucial if we want to drive individuals toward some behaviour (such as environmental friendly modes), because it might imply totally different strategies. In fact, the perfect rational theory assumes that individuals make their choices evaluating the current characteristics of
the alternatives, while bounded rationality assumes that individuals rely on other elements (previous choices, family choices and so on). Moreover, it might be important to think of different microeconomic formulations that relax some of the assumptions of perfect rationality. It seems that just relaxing the assumptions associated with the linear expansion of the utility function, already allows us to account for non-compensatory effects. More work is needed to confirm this, but it would appear that bounded effects can be accounted in the general microeconomic problem.

Acknowledgements

I would like to thank firstly Chandra Bhat and Ram Pendyala for having given me the great opportunity to prepare this resource paper. It has been a difficult but extremely interesting task and I have personally learnt a lot. I would also like to thank Juan de Dios Ortúzar, for his invaluable suggestions, comments and help during the preparation of this paper, and Italo Meloni with whom I have often engaged in interesting discussions on decision process issues. Finally, I also benefited from the discussion raised about these issues at the Discrete Choice Models Workshop organized by Michel Bierlaire in August, 2009.

References


Blase, J. (1979) Hysteresis and catastrophe theory: a demonstration of habit and threshold in travel behaviour. 7th *PTRC Summer Annual Meeting*, University of Warwick.


Elisabetta Cherchi


Yáñez, M.F., Cherchi, E. and Ortúzar J.de D. (2009a) Inertia and shock effects on mode choice panel data: implications of the transantiago implementation. 12th International Conference on Travel Behaviour Research, Jaipur, India.
