Belgian Journal of Operations Research, Statistics and Computer Science, Vol 27, n° 3

TUTORIAL PAPER XXVI

CATEGORICAL DATA ANALYSIS AN OVERVIEW OF EXPLANATORY DISCRETE DATA METHODS AND DISAGGREGATE CHOICE MODELS

Manfred FISCHER
Dept of Geography
University of Vienna
Vienna

and

Peter NIJKAMP
Dept of Economics
Free University
Amsterdam

ABSTRACT

This paper focuses attention on recent advances in the area of discrete data analysis and probabilistic choice models. The emphasis is put on the use of these methods in an explanatory context. The first part of the paper provides a survey of recently developed explanatory discrete (categorical) data methods (linear logit models, log-linear models, e.g.). The second part is concerned with explanatory discrete choice models (multinomial logit models, generalized extreme value models, e.g.). The paper concludes with a discussion of promising new research directions in the area of discrete data and choice analysis.

1. INTRODUCTION

Many data in social science research are - in contrast to those in natural sciences - measured in a discrete rather than a metric way, the major reason being that most measurement procedures in social sciences (interviews, e.g.) have only a limited degree of precision. Here the term 'discrete' measurement is used to refer to both (dichotomous and polytomous) nominal and ordinal variables. In contrast to metric variables such discrete variables can only have values in a limited set of measurement categories (see also Roberts, 1979).

Methods and models for dealing with discrete data have already a long tradition in psychometrics, sociometrics and applied statistics. In the past decade, several of those discrete date methods and models have also been applied in the area of consumer behaviour and marketing. Only in recent years, non-metric data analysis has also received profound attention in the main stream of economics. One of the fields where the application of discrete data methods and models has shown a considerable progress, is regional and urban economics, transportation economics and socio-economic geography, as many individual choice data are resulting from surveys, questionnaires or interviews. Spatial dimensions of consumer behaviour have been intensively analyzed in locational problems, migration decisions and residential choice problems.

In the present paper, a review of methods and models for analyzing choice behaviour of individuals or groups will be provided, with a particular emphasis on spatial aspects.

Generally speaking, discrete data analysis can be subdivided into two main areas, viz. exploratory discrete data analysis and explanatory discrete data analysis. Exploratory discrete analysis tends to suggest

and generate rather than to test hypotheses; it basically aims at identifying and understanding complex data structures. In contrast to explanatory discrete data analysis no (explicit) definite statistical or econometric model is assumed and tested. Explanatory analysis aims at providing insight into real-world processes or structures on the basis of priori specified testable hypotheses.

There is a wide variety of exploratory statistical procedures such as ordinal and nominal principal component analysis, factor analysis and cluster analysis, correspondence analysis, geometric and homogeneous scaling and symmetric log-linear modeling (see for further details also Bahrenberg et al., 1984, and Nijkamp et al., 1984).

The main aim of the present paper is to highlight some recent major methodological developments in the area of explanatory discrete data analysis. Explanatory analysis attempts to test the existence of causal structural relationships between endogenous end exogenous variables or between endogenous variables mutually. Explanatory methods and models can be categorized into two distinct classes, viz. explanatory discrete data analysis in a strict sense and explanatory discrete choice analysis. Discrete data analysis in a strict sense aims to analyse cause-effect relationships between a set of independent variables and one or more dependent variables, where at least the dependent variables are discrete in nature. This class will be discussed in section 2. Discrete choice analysis aims to analyse the behaviour of (groups of) individuals in a certain discrete decision context (residential migration, travel mode choice, labour force participation, e.g.). This class will be the subject of section 3. The abovementioned classes are summarized in Figure 1.

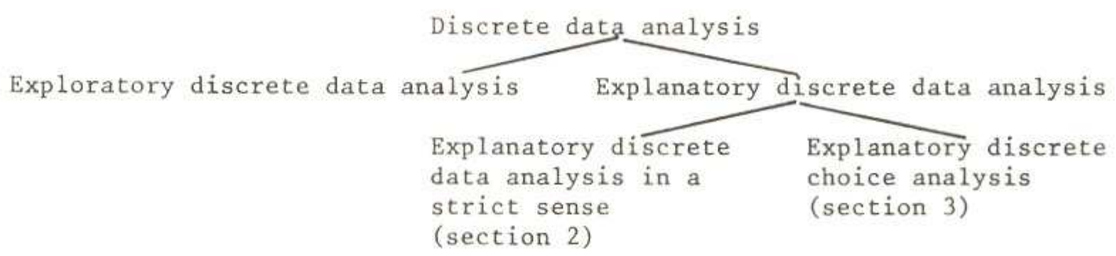

Figure 1. Categories of discrete data analysis.

It is worth noting that the distinction between explanatory discrete data analysis in a strict sense and explanatory discrete choice analysis is not always sharply demarcated. For the sake of a systematic treatment however, it is reasonable to separate these two classes (see also Fischer and Nijkamp, 1984).

2. EXPLANATORY DISCRETE DATA ANALYSIS

2.1. Introduction

Various classes of statistical/econometric methods and models in an explanatory context can be distinguished. In the framework of discrete versus metric variables Fienberg (1981) has made the following classification of explanatory methods and models based on the level of measurement of the pertaining variables (see Table 1).

		independent variables		
		discrete	metric	mixed
dependent variables	discrete	A	B	C
	metric	D	E	F

Table 1. Classes of statistical problems (Fienberg, 1981, p.3).

Conventional explanatory metric data analysis is especially aligned to cases D, E and F of Table 1. This last row of the table refers to problems which can be dealt with by means of standard multivariate procedures. For instance, in case of observable (manifest) variables, analysis of variance may be used for case D , regression analysis for case E and covariance analysis or regression analysis with dummy variables for case F . Alternatively, in case of manifest and latent variables structural equation models with latent variables based inter alia on LISREL approaches (see,for instance, Jöreskog, 1973) or PLS approaches (see, for instance, Wold, 1984) may be employed.

Explanatory discrete data analysis is focusing attention on problems of type A, B and C. These classes will be further discussed in the next subsection.

2.2. A Concise Survey of Some Explanatory Discrete Data Methods

Various kinds of explanatory discrete data methods and models may be distinguished. A main subdivision can be made according to manifest variables only and manifest and latent variables simultaneously (see also subsection 2.1.).

In the first case of manifest variables only, linear logistic regression and linear logit models have become popular tools; they may be regarded as discrete analogues of conventional regression and analysis of variance models. These models may be used for all categories A, B and C from Table 1, in as far as the variables are measured in a nominal sense. The parameters of these models are usually estimated by a maximum likelihood (ML) procedure, although in case of type A also the GSK-approach (see Grizzle et al., 1969), a non-iterative weighted least squares procedure, has turned out to be a powerful tool.

Recently, also more general types of models for manifest nominal variables have been developed, especially quantal response models which include also probit models as a special case (see Finney, 1971).
Furthermore, in a case of problems of type A also asymmetric log-linear models, a special case of the general log-linear model, can be used.
In the latter case only the marginals of a contingency table corresponding to the independent variables have to be treated as fixed, so that the product-multinomial sampling scheme with independent multinomial samples for each independent by independent variables combination has to be taken.
In this way log-linear models can be utilized to assess the effects of the independent variables upon the dependent ones. In addition, the interrelationships between the dependent variables mutually can be identified as well.

The log-linear model parameters can be estimated in various ways: by an iterative proportional fitting procedure (via ECTA or BMDP3F, e.g.), by the abovementioned GSK-approach (via NONMET II, e.g.), and by the iterative weighted least squares procedure developed by Nelder and Wedderburn (1982) (via GLIM III, e.g.).

It is worth noting that neither the linear logit model nor the log-linear model take into account an ordinal ranking that might exist among categories of dependent variables. In case of manifest ordinal dependent variables one may use the proportional odds and/or the proportional hazards model developed by McCullagh (1980). The above mentioned categories of discrete models have in recent years become a well established part of the research methodology in social sciences (including economics and geography). Less well-known are however latent class models, a set of discrete analogues of structural equation models with latent variables (see also Muthén, 1983). These models aim at analysing the relationships among a set of discrete variables some of them being manifest and others latent. There is a great diversity in latent class models. Usually the assumption is made that the manifest variables are conditionally independent given the latent variables.

In a way analogous to factor analysis two broad families of latent class models can be distinguished, viz. restricted and unrestricted models. Restricted latent class models are defined by specific patterns of *a priori* fixed values of the conditional probabilities, by equality restrictions on conditional probabilities in the same latent class, or by equality restrictions on conditional probabilities in different latent classes. ML-estimates of pertaining variables can be obtained by means of the iterative proportional fitting procedure or by Fisher's scoring procedure, a variant of the Newton-Raphson technique.

2.3. The Generalized Linear Model Approach as an Integrating Framework

Recently, much progress has been made in integrating various explanatory discrete data models into a generalized framework, the generalized linear model (GLM) approach, and in linking them to conventional linear metric data models. This unifying framework has been advocated by Nelder and Wedderburn (1974) and has been implemented in the computer package GLIM. It may encompass all categories of GLM's, as it is based on a common unifying estimation procedure (the iterative weighted least squares).

A GLM may be specified in general terms as:

$$y_i = g^{-1}(\eta_i) + \varepsilon_i \qquad i=1,\ldots,I \qquad (1)$$

with:

y_i : dependent variable stemming from an exponential family of probability distributions

ε_i : random term

η_i : linear predictor, defined as follows:

$$\eta_i = \sum_{k=1}^{K} \beta_k x_{ik} \qquad (2)$$

with:

x_{ik} : measurements of K independent variables

β_k : parameters to be estimated.

In addition, g is a (monotonic twice differentiable) link function which is defined as:

$$\eta_i = g(\mu_i) \qquad (3)$$

with the theoretical mean equal to:

$$\mu_i = E(y_i) \qquad (4)$$

In order to specify a certain GLM it is necessary to define explicitly the linear predictor, the link function and the error distribution.

The specification of the linear predictor depends mainly on the data generation process and the experimental design of the analysis. Some standard members of the GLM family characterized by a specific link function and an error distribution are presented in Table 2.

class of model	link function	error distribution
classical linear regression	identity	normal
symmetric log-linear	logarithmic	Poisson
asymmetric log-linear	logit	binomial or multi-nomial
logit regression	logit	binomial or multi-nomial
probit regression	probit	binomial or multi-nomial

Table 2. Some representative GLM's.

As far as link functions are concerned, there is a wide variety of such functions. For discrete data models, appropriate links are provided by the logit transformation (for example, in the case of the asymmetric log-linear model and the linear logit regression model), the probit transformation (for example, in the case of the probit regression model) or the logarithmic transformation (for example, in the case of the symmetric log-linear model). It should be added that also other variants of link functions are possible, for instance, the identity function in the case of linear metric data regression models.

Finally, the error distribution requires some closer attention. The exponential family of probability density function includes many continuous and discrete probability functions, such as the normal, gamma, Weibull or chi-square probability functions in the continuous

case and the Poisson, binomial, multinomial or hypergeometric probability functions in the discrete case.

It should be noted that the Poisson distribution plays a similar central role in discrete data analysis (including counted data) as the normal distribution does in metric data analysis. Generalizations of the Poisson distribution include inter alia the binomial distribution (if the dependent variable has two categories) and the multinomial distribution (if the dependent variable has multiple categories, so that this distribution may be conceived of as a constrained Poisson distribution). The error components of the asymmetric log-linear, the logit regression and the probit regression models are either defined by the binomial or by the multinomial distribution.

Recently, also some progress has been made in extending the standard GLM-approach by integrating composite link function models, quasi-likelihood models and mixture models (see also Arminger, 1984, Flowerdew and Aitkin, 1982, and Nelder, 1984).

In contrast to the simple form of a link function in the standard GLM model (based on a one-to-one relationship between μ_i and η_i-variables), composite link functions allow each μ_i to be a linear combination of some intermediate quantities (say κ_j), which are themselves functions of η_i. Examples of composite link functions can be found in latent class approaches (see also Fischer and Nijkamp, 1984), and in the proportional odds and hazards approach for ordinal variables (see McCullagh, 1980).

The class of quasi-likelihood models is marked by an incompletely defined distribution. The only assumption made is that the variance is a given function of the mean. For an example of such models in a spatial context, see Aufhauser and Fischer (1984).
Finally, mixture models take for granted that the error distribution is a mixture of several components instead of being homogeneous. Such mixtures may be continuous or discrete. In the latter case, the number of components may be known or unknown. An example of the latter approach can be found in the compound Poisson migration developed by Flowerdew and Aitkin (1982).

It may be concluded that the GLM approach offers a fruitful and unifying research area for explanatory discrete data methods and models.

2.4. Spatial Dimensions in Discrete Data Models

The use of explanatory discrete data models has exhibited some hesitating starts by the end of the seventies in the field of regional economics and geography. Although regional economists and geographers have been somewhat slow in recognizing the methodological merits of explanatory discrete data analysis for further advances in regional economic and geographical research, there are recently various signs indicating that the eighties will exhibit a major expansion of this kind of analysis in spatial research.

The major bottleneck in adopting discrete data analysis as an explanatory tool in spatial research is formed by the problem of integrating spatially or temporal-spatially dependent data in the current discrete data methodology. Thus there is an urgent need to incorporate the spatial auto- and cross-correlation research in explanatory discrete data research (see also Wrigley, 1984). A good illustration of this new direction can be found in Odland and Barff (1982) who tried to link the logic of existing space-time interaction tests to discrete data models in order to analyse the space-time patterns of American urban housing deterioration. In a spatial interaction context Fingleton (1983) has shown that considerable care is necessary when log-linear models are applied to spatially dependent data. In such cases conventional model selection procedures (for example, Brown's (1976) screening procedure and Aitkin's (1979) simultaneous test procedure) may erroneously detect interaction effects between variables which are spurious as a consequence of the spatial dependence of the measurements. Fingleton proposes to solve this problem by modifying the standard calculation of Pearson's chi-square statistic, as this statistic may take an inflated value in case of positive spatial dependence.

Fingleton (1980) has also tried to explore some of the major issues in discrete complex data sample survey designs in order to integrate spatial dependence effects in the context of log-linear modeling. Altogether, there is a continuing need to combine discrete data models with methods and models for analysing spatially dependent data.

An interesting illustration of new research directions in this framework can be found in event history analysis (see Hannan and Tuma, 1984). The main aim of event history analysis is to study discrete changes or transitions in qualitative variables. Event history analysis is based on data regarding discrete sequences and timing of transitions and it may be regarded as a potentially powerful tool in multi-period explanatory discrete data problems.

In contrast to stationary processes implied inter alia by Markov processes, event history analysis records data on all changes in a state variable within some observation period. An event history ω over a certain period $[\tau_1,\tau_2]$ can be represented as:

$$\omega[\tau_1,\tau_2] = \{y(t);\tau_1 \leq t \leq \tau_2\} \quad , \tag{5}$$

where $y(t)$ is the discrete state of a variable under consideration at period t. Each discrete 'jump' from the one episode to another one may be called an event, so that the discrete evolution of a certain phenomenon may be described by means of its event-history.

According to Hannan and Tuma (1984) there are 3 different possibilities to define statistical measures for assessing or predicting the probability of occurrence of events:

- a survivor function defining the probability that an event will occur after time t_n, given the initial information on ω_n (by means of ML methods);
- a waiting-time distribution function defining the probability of occurrence of an event based on the cumulated distribution for the waiting time (i.e., the length of intervals between successive events);
- a hazard function defining the probability of an event at time t (in terms of failing by means of a hazard rate), given that the event has not taken place before period t.

An important element in event-history analysis is not only the assessment of the probability of occurrence of an event, but - more importantly - also the assessment of which new state will be attained. Spatio-temporal applications of event-history analysis can be found inter alia in analyses

dealing with the dynamics of the labour market, of marital status and of migration patterns.

Altogether one may conclude that the study of spatial dimensions in explanatory discrete data analysis is exhibiting a promising growth path.

3. EXPLANATORY DISCRETE CHOICE ANALYSIS

3.1. Introduction

Discrete choice analysis aims to study the behaviour of (groups of) individual choice-makers based on the assumption that the set of alternative choice possibilities is finite (for example, consumer choices, migration decisions, transport mode choices, industrial locational decisions, etc.). In this respect, one may regard the frequently adopted assumption of an infinite number of alternative choices in marginalist micro-economic choice theory as rather unrealistic. Spatial choice models have recently exhibited many advances and operational applications (for instance, in the area of residential choice theory, transportation theory and labour market theory (see for instance, Domencich and McFadden, 1975, Anas, 1982, and Fischer and Maier, 1984). Such spatial models differ from general discrete choice models only with respect to the fact that the choice alternatives and/or the choice-makers are spatially distributed (though spatial interdependence effects may exist).

In the past few years an extensive body of methodological research on discrete choice analysis has been undertaken with a special emphasis on the development or use of micro-oriented utility-based models. Such models take for granted that the decision is made at the individual level based on the principle of utility maximization. The individual choice is the result of an evaluation of the expected utility associated with each discrete choice possibility implying a probabilistic choice framework.

The main theoretical underpinning of utility-based choice models can be found in Lancaster's (1971) multi-attribute utility theory and in psychological theories on individual choice behaviour (see Luce, 1959, and Tversky, 1972, e.g.).

In contrast to deterministic models, probabilistic disaggregate utility models conceive of the individual choice as a random decision. Consequently, these models may also be based on the 'bounded rationality' paradigm (see Simon, 1957), while they may also provide a meaningful analytical framework in case of unobserved or omitted relevant attributes. The next subsection will be devoted to a concise formal representation of discrete choice modeling.

3.2. A Concise Formal Introduction to Additive Random Utility Discrete Choice Models

Most additive choice random utility discrete choice models assume that the utility (u_{ia}) associated with the choice of an alternative a by an individual choice-maker i can be additively separated into two components, viz. a systematic or deterministic component, v_{ia}, and a random component, ε_{ia}. Let $A=\{1,\ldots,A'\}$ be the set of disjoint choice alternatives and $I=\{1,\ldots,I'\}$ the group of individual choice-makers. Then a random utility model can be represented as follows:

$$u_{ia} = v_{ia} + \varepsilon_{ia} \qquad (i,a) \in (I,A) \qquad (6)$$

with:

$$v_{ia} = V(z_{ia},\beta) \qquad (i,a) \in (I,A) \qquad (7)$$

and

$$\varepsilon_{ia} = \varepsilon(z_{ia},\beta) \qquad (i,a) \in (I,A) \qquad (8)$$

where z_{ia} is defined as:

$$z_{ia} := (x_i, y_a) \qquad (i,a) \in (I,A) \qquad (9)$$

while x_i is a vector of attributes characterizing choice-maker i and y_a a vector of attributes characterizing alternative a.

Next, the preference structure of choice-maker i over the relevant alternatives is defined as:

$$u_{i.} = (u_{ia}, a \in A) \tag{10}$$

Consistent choice procedures imply that alternative a will be chosen if and only if:

$$u_{ia} \geq u_{ia'} \qquad (a,a') \in (A \times A) \tag{11}$$

The deterministic component accounts for the effects of the measured alternative and the individual attributes.
The random component can represent several types of uncertainty in decision making such as imperfect information about observation, unobserved constraints that condition individual choices, unobserved attributes affecting choice, measurement errors as well as other sources of non homogeneous or inconsistent choice behaviour.

The most widely used statistical specification of the systematic component of utility is the linear-in-parameters multi-attribute model based on the theory of conjoint measurement:

$$u_{ia} = z_{ia}\beta + \varepsilon_{ia} \qquad (i,a) \in (I,A) \tag{12}$$

where the vector β being constant across individual choice-makers reflects the tastes of the individuals. Alternative specifications and extensions of the linear-in-parameters model have recently been suggested by Timmermans (1984).

The fundamental equation of random utility discrete choice models is given by the definition of the choice probabilities:

$$\begin{aligned} p(a|z_{i.},\beta) &= \text{prob}\{v(z_{ia},\beta) + \varepsilon(z_{ia},\beta) \geq \\ &\qquad v(z_{ia'},\beta) + \varepsilon(z_{ia'},\beta) \quad \bigwedge_{a' \in A} \\ &=: p_{ia} \qquad , \qquad (i,a) \in (I,A) \end{aligned} \tag{13}$$

Clearly, the choice probabilities depend on the functional form of the distribution of the error term vector. A major aim in discrete choice analysis has been to find suitable distributions that lead to computationally convenient choice probabilities and that also provide a satisfactory realistic behavioural basis.

We will first consider now the most important conventional discrete choice models which describe the behaviour of members of a group of decision-makers facing exogenously given discrete choice alternatives at a certain point in time. In this context, the hypothesis of independently and identically distributed (IID) random disturbances based on the Weibull distribution:

$$F(\varepsilon_{i.} | z_{i.}) = \prod_{a \in A} \exp\{- \exp(-\varepsilon_{ia})\} \qquad i \in I \tag{14}$$

plays a major role, as it leads directly to the family of <u>multinomial logit</u> (MNL) models. This hypothesis implies - in case of a linear-in-parameters specification - that choice-makers with identical measured attributes have identical tastes and that the correlation between unobserved or omitted alternative and individual attributes is zero.
The choice probabilities defining the MNL model can be expressed as follows:

$$p(a|z_{i.},\beta) = \exp\{v(z_{ia},\beta)\} \Big/ \sum_{a' \in A} \exp\{v(z_{ia'},\beta)\} \tag{15}$$

The MNL model is in accordance with Luce's choice axiom of <u>independence of irrelevant alternatives</u>. This so-called IIA property states that the relative choice probabilities of any two alternatives depend exclusively on their systematic utility components and are independent of other alternatives of the choice set. Due to the IIA-property, the estimation and forecasting of individual choice behaviour is considerably facilitated.

Because of its mathematical simplicity the MNL model has been preferred to other conventional discrete choice models and has been often applied in a variety of choice contexts.
It has to be recognized, however, that the underlying assumptions of the MNL model are rather restrictive and may even lead to counterintuitive behavioural predictions, especially if some alternatives are close substitutes for each other. It is not difficult to construct hypothetical

examples, such as the well-known 'red bus - blue bus' problem, that violate the IIA-property. Such violations of the premises of the MNL model may lead to inconsistent estimates of the model parameters and hence of the choice probabilities (see for an extensive discussion of MNL models also Horowitz, 1984).
Some ways to relax the restrictive assumptions of the MNL model will be discussed in the next subsection.

3.3. Generalizations of MNL Models

The computational tractability of the MNL model has caused its current popularity. At the same time, however, several attempts have been undertaken to relax the IIA-property in order to overcome the problem of similarities between alternatives. Such research efforts resulted in generalizations of the family of MNL models, *inter alia* in the class of *generalized extreme value* (GEV) models.

The latter class has been advocated by McFadden (1978) among others; it is based on the hypothesis of a broad family of multivariate extreme-value distributions of the random disturbances and it allows positive correlations among random errors, but it uses - in contrast to multinomial probit models - random terms with the same variance. This class of multivariate extreme-value distributions can be specified as:

$$F(\varepsilon_{i.} | z_{i.}) = \exp\,[-G\,\{(\exp(-\varepsilon_{ia})\ ,\ a \in A)\ ,\ z_{i.}\}] \qquad (16)$$

where G is a non-negative distribution function that is linear homogeneous in the term $(\exp(-\varepsilon_{ia})\ ,\ a \in A)$.

It can be demonstrated that the MNL model is a special case of (16). Another very interesting special type of (16) is the *nested* MNL model, which takes for granted a nested form of the decision structure, viz. the assumption that an individual choice-maker makes an initial decision independent of any other choice alternative, while in subsequent stages decisions are taken conditional to the previous one, and so forth.

Inclusive value variables representing expectations of the outcomes of lower-level decisions serve as feedback linking mechanisms of nested MNL models. In this case, the utility function shows an additive separable form. Sequential MNL models are obtained from corresponding nested MNL models when no feedback effects are incorporated into the decision structure (see, e.g. Hensher and Johnson, 1981).

Finally, it is worth noting that GEV models can also be interpreted as elimination-by-strategy (EBS) models, a general class of random preference maximizing models proposed by Tversky (1972a,b). In principle, such models allow rather general and flexible patterns of similarities between alternatives without falling into the restrictive trap of the IIA-property.

The most general family of random utility discrete choice models which circumnavigates the IIA-property can be obtained by assuming that the random disturbances are multivariate normally distributed with zero mean and an arbitrary variance-covariance matrix. Under these assumptions, the multinomial probit (MNP) choice probabilities are given by

$$p(a|z_{i'}|\beta,\Sigma_\varepsilon) = \int_{\varepsilon_{ia}=-\infty}^{\infty} \left\{ \prod_{\substack{a'=1 \\ a'\neq a}}^{A'} \int_{\varepsilon_{ia'}=-\infty}^{v(z_{ia},\beta)-v(z_{ia'},\beta)+\varepsilon_{ia}} \right\}$$

$$\left\{ N(\varepsilon_{i.}|0, \Sigma_\varepsilon)\, d\, \varepsilon_{ia'} \right\} d\, \varepsilon_{ia} \qquad (17)$$

where the number of integrals is equal to the number of alternatives and $N(\varepsilon_{i.}|0, \Sigma_\varepsilon)$ is the multivariate normal density.
MNP models have the appealing feature of allowing the random terms in the utility function to be correlated and to have unequal variances (see Daganzo, 1979). Furthermore, they also allow individual taste variation with identical observed attributes.
In contrast to the GEV models, however, the functional relationships between the choice probabilities and the measured attributes cannot be computed in an analytically closed form, except for the binary case (see (17)).

Due to the computational complexity of MNP models, they have sometimes been regarded as theoretically appealing and flexible, but practically unmanageable in case of a large number of discrete alternatives. Only quite recently, some progress has been made in providing some more effective and accurate estimation procedures, such as direct numerical integration methods (see Hausman and Wise, 1978), iterative approximation methods based on Clark's approximation (see Daganzo et al., 1977),simulated frequency methods (see Lerman and Manski, 1981), or separated split models (see Langdon, 1984). Clearly, once the computational problem of MNP models have been tackled, they may provide powerful operational tools for discrete choice analysis (see Van Lierop and Nijkamp, 1984).

3.4. New Approaches to Discrete Choice Analysis

New approaches to conventional discrete choice models are especially aligned to a multi-period context. It is clear that many distinct choices (labour force participation, residential locational decisions, e.g.) are not unique, but recurrent. In such cases, a multi-period or - preferably - a dynamic choice model has to be employed. The past few years have exhibited an increasing interest in the development of discrete choice models which explicitly incorporate dynamic aspects of choice behaviour.

A first interesting research direction can be found in stochastic panel data discrete choice approaches (see Heckman, 1981a, 1981b, Halperin, 1984, and Fischer and Nijkamp, 1984). Such approaches have been developed for analysing the structure of discrete choices in a multi-period framework and represent a more genuine behavioural research direction. In such models two main effects may be distinguished, viz. serial correlation effects and state dependence effects. Serial correlation (also termed spurious state dependence) in the observed attributes is assumed to be known to the individual choice-maker, but unknown to the analyst. State dependence results from (sequential or time-dependent) impacts on the individual's current choices from previous ones. Serial correlation is due to omitted and/or unmeasured attributes which do not (or only marginally) change over time. In his pioneering work, Heckman (1981b) illustrates that serial correlation and state dependence can be represented within the framework of a more general panel data discrete choice model.

This model includes serial correlation models, state dependence models as well as any combination of serial correlation and state dependence as specific cases. It may be added that recent advances in the field of activity-based choice analysis or longitudinal discrete choice analysis may also significantly contribute to a further development of explanatory multi-period discrete choice models (see for more details, Coleman, 1981, Halperin, 1984, and Koppelman and Pas, 1984).
The same holds true for recently applied methods of stated preference techniques in micro choice analysis (see Kroes and Sheldon, 1984).

Another interesting research development concerns attempts to endogenize the set of relevant choice alternatives. Conventional models assume that the information about alternatives is given for the choice-maker. In many cases, however, this exogeneity assumption is not very realistic, as information on alternatives may be imperfect or depend on (one's own or others') past choices. Personal experience, learning and communication may thus be important ingredients in discrete choice analysis. An interesting illustration of these ideas can be found in De Palma and Lefevre (1982), who deal with a choice context in which choice-makers interact in their decision process, so that the attributes characterizing the choice problem of a decision-maker also depend on the behaviour of other choice-makers. The authors then propose the use of a continuous-time Markov model allowing individuals to interact with others during their choice process.

New directions can also be observed in the field of non-linear choice models, notably non-linearities in the observed individual or attribute specification function. Some methods of representing variable transformation involve inter alia the Box-Cox or the Box-Tukey generalization. In the latter case the systematic utility component has a polynomial form (see Gaudry and Wills, 1978, and Longley, 1984).

Finally, new research is also being undertaken in the area of dynamic micro choice behaviour. An interesting contribution in this field was made by Ben-Akiva and De Palma (1984), who developed a model that is able to predict the decision of an individual to change his present state in two stages: the decision to change (or to undertake a transaction),

and - conditional to a change taking place - the choice of a new alternative. The specification of this choice model is based on disaggregate dynamic logit models. Altogether - in combination with activity-based and longitudinal approaches - such dynamic models may lead to a real break-through in disaggregate choice analysis.

4. ITEMS OF A RESEARCH AGENDA

The field of discrete data and choice analysis appears to provide an extremely rich research area. Despite significant advances however, there are still various open methodological problems left. A sample of such problems making up a set of items on a research agenda will be discussed here.

- There is a clear need for more appropriate estimation procedures in the context of more complex types of sampling processes including a stratification in exogenous and endogenous variables at the same time.
- More attention should be focused on estimation procedures for panel and longitudinal data discrete choice models including serial correlation and state dependence effects.
- A closer analysis and identification of spatial auto- and cross-correlation in the case of discrete spatially dependent data is necessary.
- There is a need for developing methods for forecasting aggregate population behaviour, given an estimated individual choice model and a description of the environment in which future choices are likely to take place.
- The validity and nature of discrete choice analysis in case of forecasting models deserve a critical evaluation (by means of back-casting, e.g.).
- There is much scope for the development of simultaneous equation discrete choice models in which one or more of the attributes affecting choices are dealt with endogenously.

REFERENCES

Aitkin, M., 'A simultaneous test procedure for contingency table models', Applied Statistics, vol. 28, 1979, pp. 233-242.

Anas, A., Residential Location Markets and Urban Transportation, Academic Press, New York, 1982.

Arminger, G., 'Analysis of qualitative individual data of latent class models with generalized linear models', in Nijkamp, P., Leitner, H., and Wrigley, N. (eds.), Measuring the Unmeasurable, Martinus Nijhoff, The Hague, 1984.

Aufhauser, E., and Fischer, M.M., 'Loglinear modelling and spatial analysis', paper presented at the 24th European Regional Science Conference', Milan,1984.

Bahrenberg, G., M.M. Fischer , and P. Nijkamp, (eds.), Recent Developments in Spatial Data Analysis: Methodology, Measurement, Models, Gower, Aldershot, 1984.

Ben-Akiva, M., and A. de Palma, Analysis of a Dynamic Location Choice Model with Transaction Costs, Research Paper Centre of Operations Research & Econometrics, University of Loyvain, Louvain, 1984 (mimeographed).

Brown, M.B., 'Screening effects in multidimensional contingency tables', Applied Statistics, vol. 25, 1976, pp. 37-46.

Coleman, J.S., Longitudinal Data Analysis, Basic Books, New York, 1981.

Daganzo, C., F. Bouthelier, and Y. Sheffi, 'Multinomial probit and qualitative choice: A computationally efficient algorithm', Transportation Science, vol. 11, 1977, pp. 338-358.

Daganzo, C., Multinomial Probit. The Theory and its Application to Demand Forecasting, Academic Press, New York, 1979.

De Palma, A., and C. Lefevre, 'Individual decision-making in dynamic collective systems', Paper presented at the IIASA-workshop on Spatial Choice Models in Housing, Transportation and Land Use Analysis: Towards a Unifying Effort, Laxenburg, 29th March-1st April, 1982.

Domencich, T.A., and D. McFadden, Urban Travel Demand: A Behavioral Analysis, North Holland Publ. Co., Amsterdam, 1975.

Fienberg, S.R., The Analysis of Cross-Classified Categorical Data, 2nd edition, MIT Press, Cambridge, Mass., 1981.

Fingleton, B., 'Log-linear modelling of geographical contingency tables', Environment and Planning A, vol. 13, 1981, pp. 1539-1551.

Fingleton, B., 'Log-linear models with dependent spatial data', Environment and Planning A, vol. 15, 1983, pp. 801-813.

Finney, D.J., Probit Analysis, 3rd edition, Cambridge University Press, London, 1971.

Fischer, M.M., and G. Maier, 'Discrete choice and labour supply modelling', Paper presented at the Besancon Symposium of the IGU Working Group on Systems Analysis and Mathematical Models, August 21-23, 1984.

Fischer, M.M., and P. Nijkamp, 'Categorical data and choice analysis in a spatial context', Urban Modelling, (B. Hutchinson, M. Batty, and P. Nijkamp, eds.) , Springer, Berlin, 1984.

Flowerdew, R., and M. Aitkin, 'A method for fitting the gravity model based on the Poisson distribution', Journal of Regional Science , vol. 22, 1982, pp. 191-202.

Gaudrey, M.J.I., and M.J. Wills, Estimating the functional form of travel demand models, Transportation Research, vol. 12, 1978, pp. 257-289.

Grizzle, J.E., C.F. Starmer, and G.G. Koch, 'Analysis of categorical data by linear models', Biometrics, vol. 25, 1969, pp. 489-504.

Halperin, W.C., 'The analysis of panel data for discrete choices', in: Nijkamp, P., Leitner, H., and Wrigley, N. (eds.), Measuring the Unmeasurable, Martinus Nijhoff, The Hague, 1984.

Hannan, M.T., and N.B. Tuma, 'Dynamic analysis of qualitative variables: Applications to organizational demography', in: Nijkamp, P., Leitner, H., and Wrigley, N. (eds.), Measuring the Unmeasurable, Martinus Nijhoff, The Hague, 1984.

Hausman, J.A., and Wise, D.A., 'A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences', Econometrica, vol. 46, 1978, pp. 403-426.

Heckman, J.J., 'Statistical models for discrete panel data', in: Manski, C.F., and McFadden, D. (eds.), Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, Mass., 1981a, pp. 114-178.

Heckman, J.J., 'The incidental parameters problems and the problem of initial conditions in estimating a discrete time-discrete data stochastic process', in: Manski, C.F., and McFadden, D. (eds.), Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, Mass., 1981b, pp. 179-1975.

Horowitz, J., 'Random utility models as practical tools of travel demand analysis', in: Jansen, G.R.M., Nijkamp, P., and Ruijgrok, C.J. (eds.), Transportation and Mobility in an Era of Transition, 1984.

Jöreskog, K.G., 'A general method for estimating a linear structural equation system', Structural Equation Models in the Social Sciences (A.S. Goldberger and O.P. Duncan, eds.), Seminar Press, New York, 1973, pp. 85-112.

Koppelman, F.S., and E.I. Pas, 'Travel-activity behaviour in time and space: methods for representation and analysis', Measuring the Unmeasurable, (P. Nijkamp, H. Leitner, and N. Wrigley, eds.), Martinus Nijhoff, The Hague, 1984.

Kroes, E., and R.J. Sheldon, 'The use of stated preference techniques to derive travel elasticities', in: G.R.M. Jansen, P. Nijkamp, and C. Ruijgrok (eds.), Transportation and Mobility in an Era of Transition, North-Holland Publ. Co., Amsterdam, 1984.

Lancaster, K., Consumer Demand; a New Approach, Columbia University Press, New York, 1971.

Langdon, M.G., 'Methods of determining choice probability in utility maximising multiple alternative models', Transportation Research B, vol. 18, no. 3, 1984, pp. 209-234.

Lerman, S.R., and C.F. Manski, 'On the use of simulated frequencies to approximate choice probabilities', in: Manski, C.F., and McFadden, D., (eds.), Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, Mass., 1981, pp. 305-319.

Lierop, W.F.J. van, and P. Nijkamp, 'Perspectives of disaggregate choice models on the housing market', in: Pitfield, D.A. (ed.), Discrete Choice Model in Regional Science, Pion, London, 1984, pp. 141-162.

Longley, P.A., 'Discrete choice modelling and complex spatial choice: An overview', in Bahrenberg, G., Fischer, M.M., and Nijkamp, P. (eds.), Recent Developments in Spatial Data Analysis: Methodology, Measurement, Models, Gower, Aldershot, 1984, pp. 375-391.

Luce, R.D., Individual Choice Behavior, Wiley, New York, 1959.

McCullagh, P., 'Regression models for ordinal data', Journal of the Royal Statistical Society, vol. 42B, 1980, pp. 109-142.

McFadden, D., 'Modelling the choice of residential location', in: Karlqvist, A., Lundqvist, L., Snickars, F., and Weibull, J.W. (eds.), Spatial Interaction Theory and Planning Models, North-Holland, Amsterdam, 1978, pp. 75-96.

Muthén, B., 'Latent variable structural equation modeling with categorical data', Journal of Econometrics, vol. 21, 1983, pp. 43-65.

Nelder, J.A., 'Statistical models for qualitative data', Measuring the Unmeasurable, (P. Nijkamp, H. Leitner and N. Wrigley, eds.), Martinus Nijhoff, The Hague, 1984.

Nelder, J.A., and R.W.M. Wedderburn, 'General Linear Models', Journal of the Royal Statistical Society A, 135, 1972, pp. 370-384.

Nijkamp, P., H. Leitner, and N. Wrigley (eds.), Measuring the Unmeasurable, Martinus Nijhoff, The Hague, 1984.

Odland, J., and R. Barff, 'A statistical model for the development of spatial problems: Applications to the spread of housing deterioration', Geographical Analysis, vol. 14, 1982, pp. 327-339.

Roberts, F.S., Measurement Theory with Applications to Decisionmaking, Addison-Wesley, Reading, Mass., 1979.

Simon, H.A., Models of Man, Wiley, New York, 1957.

Timmermans, H.J.P., 'Decision models for predicting preferences among multiattribute choice alternatives', in: Bahrenberg, G., Fischer, M.M., and Nijkamp, P. (eds.), Recent Developments in Spatial Data Analysis: Methodology, Measurement, Models, Gower, Aldershot, 1984, pp. 337-354.

Tversky, A., 'Elimination-by-aspects: A theory of choice', Psychological Review, vol. 79, 1972a, pp. 281-299.

Tversky, A., 'Choice-by-elimination', Journal of Mathematical Psychology, vol. 9, 1972b, pp. 341-367.

Wold, H., 'Systems analysis by partial least squares', Measuring the Unmeasurable (P. Nijkamp, H. Leitner, and N. Wrigley, eds.), Martinus Nijhoff, The Hague, 1984.

Wrigley, N., 'Categorical data methods and discrete choice modelling in spatial analysis: Some directions for the 1980's', in: Nijkamp, P., Leitner. H., and Wrigley, N. (eds.), Measuring the Unmeasurable, Martinus Nijhoff, The Hague, 1984.

INSTRUCTIONS FOR THE AUTHORS

1. JORBEL accepts papers in the fields of Operations Research, Statistics and Computer Science.

2. Theoretical, applied and didactical papers, as well as documented computer programs are considered.

3. As far as possible, each issue will include a TUTORIAL PAPER. Authors are invited to submit also such papers giving didactical surveys on specific themes. These papers can be theoretical or applied. They should be comprehensible for non specialists and should establish the link between research and practice.

4. For the sake of effective communication it is recommended to the authors to have their papers written in *English.*

5. All submitted papers will be refereed. Only the authors will be responsible for their text.

6. As the journal is printed by offset, the manuscript should be camera ready, including figures and tables.

The *first page* is composed by the printer. It should only include:
- The title of the paper
- The names and affiliations of the authors (address included)
- An abstract of no more than 8 lines in English.

It is recommended to have the text of the manuscript typed double spaced with a prestige Elite type-head and the paragraph headings with a Script.
The formulas should also be typed and numbered on the right-hand side.
At each paragraph an 8 character jump should be observed.

7. The first author of published papers will receive 25 copies.

8. Papers should be submitted in two copies to one of the principal editors.

The authors should mention whether it is a tutorial paper.

ANNUAL SUBSCRIPTION RATE (4 issues)
BELGIUM: 800 BF. OTHER COUNTRIES: 900 BF

Sogesci: 53, rue de la Concorde, B-1050 Bruxelles, Belgium
B.V.W.B.: Eendrachtstraat 53, B-1050 Brussel, Belgium

Bank Account: 000-0027041-75

SOGESCI - B.V.W.B. BULLETIN

A Bulletin giving information on Operations Research, Statistics and Computer Science events is published quarterly by the SOGESCI-B.V.W.B.

Information on such events are to be mailed to G. Janssens:
RUCA, Middelheimlaan 1, B-2020 Antwerpen, Belgium