Transpn Res.-B Vol. 17B, No. 1, pp. 13–23, 1983
Printed in Great Britain.

0191-2615/83/010013-11$03.00/0

DISCRETE CHOICE THEORY, INFORMATION THEORY AND THE MULTINOMIAL LOGIT AND GRAVITY MODELS

ALEX ANAS
Department of Civil Engineering, Northwestern University, Evanston, IL 60201, U.S.A.

(*Received* 18 *May* 1981)

Abstract—The strong "similarity" between "information minimizing" and "utility maximizing" models of spatial interaction has been known for some time (see Anas 1975, Williams, 1977), but the extent of this "similarity" has been underestimated. This paper proves that the two approaches are identical in that the multinomial logit model can be derived and identically estimated by either method. It is also proved that the doubly-constrained gravity model derived by Wilson (1967) is identical to a multinomial logit model of joint origin-destination choice, consistent with stochastic utility maximization. It follows that behaviorally valid "gravity models" can be estimated from disaggregated data on individual choices. In closure, "behavioral demand modeling", which follows McFadden (1973), and "entropy-maximizing modeling", which follows Wilson (1967), should be seen as two equivalent views of the same problem. The behavioral content of models estimated by either approach is entirely determined by the model specification and data aggregation beliefs of the analysts, and not by any inherent structural property of the models themselves.

Gravity models have a long history as tools invented to help the transportation planner balance an origin-to-destination trip table so that the predicted zone-to-zone flows are consistent with the trips generated at each origin and the trips terminating at each destination. The crudest gravity model has the form,

$$N_{ij} = G_{ij} \frac{O_i D_j}{f(d_{ij})} \tag{1}$$

where O_i are the trips originating from zone i; D_j are the trips terminating in zone j; $f(d_{ij})$ is the impedance (or separation) between i and j as an increasing function of the distance from i to j; N_{ij} is the predicted trips originating from zone i and terminating in zone j; and G_{ij} is an "adjustment factor", the value of which is selected to "balance" trips by assuring that $\sum_i N_{ij} = D_j$ and $\sum_j N_{ij} = O_i$ for each i and j.

Gravity models became widely used and abused in the fifties and sixties, and it was not until 1967 that Wilson provided the first theoretically valid derivation of the gravity model from statistical information-minimizing (or entropy-maximizing) principles. Wilson's work brought elegance, analytical rigor with a long-awaited sense of closure to the raging confusion about the interpretability and theoretical integrity of the gravity model. Following Wilson's work the application of entropy-maximizing models flourished in Britain and to a lesser extent, in the U.S. and elsewhere. Transportation modeling became enhanced by a new generation of gravity models. Many practicing transportation planners were able to correct their previous abuses of the gravity model and to balance their trip tables with a new sense of statistical consistency.

While these developments were occurring, McFadden (1973), Ben-Akiva (1973) and others began formulating the problems of travel mode and location decisions as problems in micro-economic consumer choice among discrete alternatives. McFadden's work was preceded by a decade of empirical research stemming from Warner's (1962) binary choice analysis. Much like the empirical research on gravity models, the work on choice modeling preceding McFadden's contributions had not succeeded in providing a theoretical grounding of empirically established concepts. McFadden's derivation of the logit model from utility maximization closed this gap and a new area of research, labeled "behavioral demand modeling", emerged. This research area stressed the importance of stochastic utility maximization and the use of disaggregate small-sample data in the estimation of choice models via maximum likelihood.

The strong similarities between the multinomial logit model of behavioral modeling and the spatial interaction model of entropy maximization did not remain unnoticed. Syntheses of entropy-maximization and utility-maximization concepts in different contexts have been pro-

posed by Anas (1975), Williams (1977) and Los (1979). Despite the search for a conclusive synthesis, the two approaches remained apart because their perceived similarities were greatly obscured by important differences in prevailing practices in the use of data. Analysts working within Wilson's entropy-maximization paradigm estimated models from aggregated data, while behavioral demand modelers shifted their emphasis to small-sample disaggregate data.

This paper proves that the minimum-information and the behavioral discrete choice modelling approaches are identical. The different estimation methods of the two approaches applied to the same data yield identical coefficients for the multinomial logit model and produce identical predictions of actual choice patterns. The two paradigms imply mutually consistent and fully equivalent model search and model specification strategies: one is rooted in micro-behavioral postulates, the other in macro-statistical information theory.

It is also shown in this paper that the traditional spatial gravity model (1), rederived by Wilson (1967) as an entropy model, is identical to a multinomial logit model of joint origin-destination choice derived from stochastic utility maximization.

1. STOCHASTIC UTILITY MAXIMIZATION AND THE MULTINOMIAL LOGIT MODEL

The work of McFadden (1973) and others has accomplished two important tasks: (1) the theoretical grounding of discrete choice models in utility maximization, and (2) the establishment of the econometrics of maximum likelihood as the estimation method for choice models.

By far the most tractable and widely used of available choice models is the multinomial logit (MNL) model. We now briefly review the theoretical derivation of the MNL model.

Consider a population of $h = 1 \ldots H$ individual decision-makers (hereafter, *choosers*) who have homogeneous preferences up to an additive stochastic term. Each chooser faces a choice among $j = 1 \ldots J$ discrete alternatives. The set of these J alternatives is called the choice set. The utility of each alternative j is assumed to be a linear function of the utility attributes or of predetermined non-linear transformations of these attributes. Hence,

$$\hat{U}_j^h = \alpha_{oj} + \sum_{k=1}^{K} \alpha_k X_{jk}^h + \epsilon_j^h \tag{2}$$

where $\hat{U}_j^h$ is the perceived utility of alternative j for chooser h; $\bar{\alpha} = [\alpha_{o1}\alpha_{o2} \ldots \alpha_{oJ}\alpha_1\alpha_2 \ldots \alpha_K]$ are the utility coefficients common to all choosers in the population; X_{jk}^h is the value of the kth attribute (or transformed attribute) for alternative j and chooser h; and $\bar{\epsilon} = [\epsilon_1\epsilon_2 \ldots \epsilon_J]$ is the vector of stochastic utility which is distributed over the population. The coefficients α_{0j} are alternative-specific constants, because they measure the unspecified part of utility for each alternative.

The utility-maximizing choice model is derived from

$$P_i^h = \text{Prob.}\,[\hat{U}_i^h > \hat{U}_j^h;\ \forall j \neq i] \tag{3}$$

where P_i^h is the probability that chooser h chooses alternative i. The MNL model is derived by assuming that each ϵ_i^h is independently and identically distributed (IID) over the population and for each chooser according to the Gumbel distribution which has the cumulative distribution function,

$$\text{Prob.}\,(\epsilon_i^h \leqq \epsilon) = \exp\left(-\exp\left[-\left(\frac{\pi^2}{6\sigma^2}\right)^{1/2}\epsilon\right]\right) \tag{4}$$

with mode zero and variance σ^2 for each alternative $i = 1 \ldots J$. The MNL model, thus derived, has the form

$$P_i^h = \frac{\exp\left\{\beta_{oi} + \sum_{k=1}^{K} \beta_k X_{ik}^h\right\}}{\sum_{j=1}^{J} \exp\left\{\beta_{oj} + \sum_{k=1}^{K} \beta_k X_{jk}^h\right\}} \tag{5}$$

where $\beta_{oi} = (\pi^2/6\sigma^2)^{1/2}\alpha_{oi}$ and $\beta_k = (\pi^2/6\sigma^2)^{1/2}\alpha_k$. The utility coefficients $\bar{\alpha}$ cannot be identified, but the scaled coefficients $\bar{\beta} = [\beta_{o1} \ldots \beta_{oJ}\beta_1 \ldots \beta_K]$ are uniquely estimable with the exception of one of the alternative-specific constants, say β_{oJ}. The differences $\beta_{oj} - \beta_{oJ}$ are uniquely estimable for each j.

To estimate the model we maximize the likelihood function (joint probability that the observed choices are generated by the model) with respect to the estimable coefficients $\bar{\beta}$. Thus,

$$\underset{\bar{\beta}}{\text{Maximize}} \log \mathscr{L} = \sum_h \sum_j \delta_j^h \log P_j^h(\bar{\beta}) \tag{6}$$

where $\delta_j^h = 1$ if chooser h chooses alternative j and $\delta_j^h = 0$ if chooser h does not choose alternative j. The first-order equations for this unconstrained maximization problem are

$$\frac{\partial \log \mathscr{L}}{\partial \bar{\beta}} = \sum_h \sum_j \delta_j^h \left[\frac{\partial P_j^h(\bar{\beta})/\partial \bar{\beta}}{P_j^h(\bar{\beta})}\right] = 0 \tag{7}$$

where

$$\frac{\partial P_j^h(\bar{\beta})}{\partial \beta_k} = P_j^h(\bar{\beta})X_{jk}^h - P_j^h(\bar{\beta})\sum_i P_i^h(\bar{\beta})X_{ik}^h; \qquad k = 1 \ldots K \tag{8}$$

$$\frac{\partial P_j^h(\bar{\beta})}{\partial \beta_{oj}} = P_j^h(\bar{\beta})[1 - P_j^h(\bar{\beta})]; \qquad j = 1 \ldots J. \tag{9}$$

Substituting (8) and (9) into (7):

$$\frac{\partial \log \mathscr{L}}{\partial \beta_k} = \sum_h \sum_j P_j^h(\bar{\beta})X_{jk}^h - \sum_h \sum_j \delta_j^h X_{jk}^h = 0; \qquad k = 1 \ldots K \tag{10}$$

and

$$\frac{\partial \log \mathscr{L}}{\partial \beta_{oj}} = \sum_h P_j^h(\bar{\beta}) - \sum_h \delta_j^h = 0; \qquad j = 1 \ldots J. \tag{11}$$

where $\sum_h \sum_j \delta_j^h X_{jk}^h \equiv \bar{X}_k$ is the aggregate value of the kth attribute over the estimation data, and $\sum_h \delta_j^h \equiv \mathring{N}_j$ are the observed frequencies of choosers choosing each of the j alternatives. The ratio $\mathring{N}_j/H$ is also called the *market share* of the jth alternative.

Solving the $K + J$ eqns in (10) and (11) simultaneously, we obtain all the elements of $\bar{\beta}$; and this is an MNL model with K generic attributes and a full set of alternative-specific constants, all but one of which are identified. It follows from (11) that the estimated alternative-specific constants insure perfect predictions of the market shares and it follows from (10) that the generic-attribute coefficients insure perfect predictions of the mean value of each generic attribute.

The assumptions needed to derive the MNL model are quite strong, and can be summarized as follows:

(A1) All choosers in the population have the same utility function, which is linear in attributes or linear in predetermined transformations of the attributes (such as loglinear, quadratic, etc.).

(A2) The utility coefficients do not vary over the population of choosers (i.e. there is no taste variation).

(A3) The stochastic part of utility is additive and is Gumbel distributed in the population of choosers and for each chooser, with mode zero and variance σ^2 for each alternative.

(A4) Each chooser maximizes utility, thus choosing the most preferred alternative.

2. INFORMATION MINIMIZATION AND THE MULTINOMIAL LOGIT MODEL

Within Wilson's information-minimization (entropy-maximization) approach, models are not derived from micro-behavioral postulates but from information-theoretic principles which seek

to find the most random predictions of individual choices consistent with observations on the aggregate (macro) or average (mean) states of the entire population of choosers.

We shall now state the information-minimization problem which fully corresponds to the maximum-likelihood problem given by (5). This is obtained by minimizing Shannon's measure of the information in a probability distribution (or minus-one times the entropy $\mathscr{E}$), subject to constraints on the aggregate values of the attributes and choices. Thus,

$$\underset{[P_j^h]}{\text{Minimize}} -\mathscr{E} = \sum_h \sum_j P_j^h \log P_j^h \tag{12}$$

subject to:

$$\sum_j P_j^h = 1; \qquad h = 1 \ldots H \tag{13}$$

$$\sum_h P_j^h = \sum_h \delta_j^h; \qquad j = 1 \ldots J \tag{14}$$

$$\sum_h \sum_j P_j^h X_{jk}^h = \sum_h \sum_j \delta_j^h X_{jk}^h; \qquad k = 1 \ldots K. \tag{15}$$

In this formulation the unknowns are the choice probabilities $[P_j^h]$. We seek the most random (information-minimizing) predictions $[P_j^h]$, but we require that these predictions replicate the aggregate observations on the entire system. These requirements are imposed in the constraints. Constraint (14) states that the predicted expectation of choosers choosing each alternative should equal the actual number of choosers choosing it. Constraint (15) states that the expectation of the aggregate value of each attribute should equal the observed aggregate value. Constraint (13) states that the choice probabilities for each chooser should sum to unity. Because the objective function is convex and the constraints linear, the solution of this optimization problem generates unique probabilities $[P_j^h]$.

If we minimize (12) subject only to (13), we obtain the result that $P_j^h = 1/J$ for each j and each h: if there is no structure imposed by observation of the aggregates of the system, then the best (most knowledgeable) prediction is an entirely random one, i.e. the uniform distribution or equiprobable assignment of choosers to alternatives. By imposing constraints (15), (14) and (13) we obtain a model which replicates all market shares and all generic attribute aggregate values. The probabilities obtained are always the most random possible, but will replicate the required macro information.

To obtain the analytical solution we form the Lagrangian of (12)–(15). This is

$$L_{\mathscr{E}} = \sum_h \sum_j P_j^h \log P_j^h - \sum_h \theta_h \left[\sum_j P_j^h - 1\right] - \sum_j \lambda_{oj} \left[\sum_h P_j^h - \mathring{N}_j\right] - \sum_k \lambda_k \left[\sum_h \sum_j {}_j P_j^h X_{jk}^h - \bar{X}_k\right] \tag{16}$$

where θ_h, λ_{oj} and λ_k are the Lagrange multipliers and $\mathring{N}_j = \sum_h \delta_j^h$, $\bar{X}_k = \sum_j \sum_h \delta_j^h X_{jk}^h$. The first-order conditions are

$$\frac{\partial L_{\mathscr{E}}}{\partial P_j^h} = 1 + \log P_j^h - \theta_h - \lambda_{oj} - \sum_k \lambda_k X_{jk}^h = 0; \qquad j = 1 \ldots J, \quad h = 1 \ldots H \tag{17}$$

$$\frac{\partial L_{\mathscr{E}}}{\partial \theta_h} = \sum_j P_j^h - 1 = 0; \qquad h = 1 \ldots H, \tag{18}$$

$$\frac{\partial L_{\mathscr{E}}}{\partial \lambda_{oj}} = \sum_h P_j^h - \sum_h \delta_j^h = 0; \qquad j - 1 \ldots J, \tag{19}$$

$$\frac{\partial L_{\mathscr{E}}}{\partial \lambda_k} = \sum_h \sum_j P_j^h X_{jk}^h - \sum_j \sum_h \delta_j^h X_{jk}^h = 0; \qquad k = 1 \ldots K. \tag{20}$$

From (17),

$$P_j^h = \exp(-1 + \theta_h + \lambda_{oj} + \sum_k \lambda_k X_{jk}^h) \quad \text{or,}$$

$$P_j^h = \exp(-1 + \theta_h) \exp\left(\lambda_{oj} + \sum_k \lambda_k X_{jk}^h\right). \tag{21}$$

Substituting (21) into (18) we find,

$$\exp(-1 + \theta_h) = 1 \Big/ \sum_j \exp\left(\lambda_{oj} + \sum_k \lambda_k X_{jk}^h\right) \tag{22}$$

and substituting this into (21), we obtain

$$P_i^h = \frac{\exp\left\{\lambda_{oi} + \sum_k \lambda_k X_{ik}^h\right\}}{\sum_j \exp\left\{\lambda_{oj} + \sum_k \lambda_k X_{jk}^h\right\}} \tag{23}$$

which is the multinomial logit model identical in form to (5). The Lagrange multipliers θ_h, $h = 1 \ldots H$ have been eliminated in insuring that (23) satisfies (18). To estimate λ_{oj}, $j = 1 \ldots J$ and λ_k, $k = 1 \ldots K$, we substitute (23) into (19) and (20) and solve the resulting $J + K$ equations for the values of these Lagrange multipliers.

Letting $\vec{\lambda} = [\lambda_{o1}\lambda_{o2} \ldots \lambda_{oJ}\lambda_1 \ldots \lambda_K]$, eqns (20) and (19) which must be solved to find $\vec{\lambda}$, can be written as,

$$\frac{\partial L_{\vec{\lambda}}}{\partial \lambda_k} = \sum_h \sum_j P_j^h(\vec{\lambda}) X_{jk}^h - \sum_h \sum_j \delta_j^h X_{jk}^h = 0; \qquad k = 1 \ldots K \tag{24}$$

$$\frac{\partial L_{\vec{\lambda}}}{\partial \lambda_{oj}} = \sum_h P_j^h(\vec{\lambda}) - \sum_h \delta_j^h = 0; \qquad j = 1 \ldots J. \tag{25}$$

The assumptions needed to derive the MNL model through information minimization are quite weak. In summary, these assumptions are as follows:

(A4): The most probable prediction of the choice probabilities is that which minimizes the information in these probabilities (maximizes randomness, entropy) subject to available information.

(A5): The predicted choice probabilities must replicate certain macro properties of the aggregate system of choosers: (1) the predicted expected number of choosers choosing each alternative should equal the number of actual choosers choosing that alternative; (2) the expected total value of each attribute should equal the observed total value of that attribute.

Our stated information-minimizing derivation differs from Wilson's original derivation only because we have presented ours for the case of individual (chooser-specific) choice probabilities. Wilson's original formulation was based on relative frequencies because he dealt with aggregations of choosers, each aggregation corresponding to a spatial zone. Wilson's original formulation will be considered in Section 5, while the problem of aggregation is reviewed in Section 4.

3. PROOF THAT LIKELIHOOD MAXIMIZATION AND INFORMATION MINIMIZATION YIELD IDENTICAL ESTIMATES FOR THE COEFFICIENTS OF THE MNL MODEL

We now prove that the MNL model can be identically estimated via maximum likelihood, which finds the utility coefficients, $\vec{\beta}$, or via information minimization, which finds the Lagrange multipliers, $\vec{\lambda}$.

Theorem: Let $\vec{\beta}^*$ be the maximum likelihood (ML) estimate of $\vec{\beta}$ obtained via (6), and let $\vec{\lambda}^*$ be the minimum information (MI) Lagrange multipliers of (12) subject to (13)–(15). Given the

same data $[\delta_j^h]$, $[X_{jk}^h]$ and $[\bar{X}_k]$, the two problems (i.e. (6) and (12) subject to (13)–(15)), have identical solutions, i.e. $\bar{\beta}^* = \bar{\lambda}^*$.

Proof: It has been shown that ML and MI can be used interchangeably to derive the MNL model (5) or (23). To find $\bar{\beta}^*$ one solves (10) and (11), and to find $\bar{\lambda}^*$ one solves (24) and (25). Since these two sets of equations are identical and since $P_j^h(\bar{\beta})$ and $P_j^h(\bar{\lambda})$ are the same function, it follows that $\bar{\beta}^* = \bar{\lambda}^*$. QED.

Corollary 1: To estimate correctly the coefficients of the MNL model, the minimum amount of observed data are the following: (i) the number of choosers choosing each alternative ($\mathring{N}_j = \sum_h \delta_j^h$, $j = 1 \ldots J$); (ii) the level of each generic attribute k for each chooser h and each choice alternative j (X_{jk}^h, $h = 1 \ldots H$, $j = 1 \ldots J$, $k = 1 \ldots K$); (iii) the aggregate level of each generic attribute ($\bar{X}_k$, $k = 1 \ldots K$). If the aggregate level of each generic attribute and the aggregate choices are directly observed, then the matrix of individual choices, $[\delta_j^h]$, need not be observed.

Proof: The proof follows directly by examining the form of the eqns (10) and (11) or (24) and (25). It can be seen from these equations that observing individual choices $[\delta_j^h]$ is not necessary if the left sides of $\bar{X}_k = \sum_h \sum_j \delta_j^h X_{jk}^h$ and $\mathring{N}_j = \sum_h \delta_j^h$ can be observed directly without observing the choice of each chooser. Indeed, the only reason ever to have to observe the actual choice of each chooser is that this information is used solely to compute the aggregates $\bar{X}_k$ and $\mathring{N}_j$, and makes no other contribution to the estimation process. QED.

This corollary has an important implication for survey and questionnaire design in "disaggregate behavioral demand modelling". Such surveys are necessary only in order to obtain the attribute information $[X_{jk}^h]$, but can also be extended at negligible marginal cost to obtain data on actual choices, $[\delta_j^h]$. It is interesting to note, however, that if the *aggregates* of attributes such as travel time, travel cost, housing price, etc. can be estimated from independent sources, then the need to know the actual choices of choosers disappears: so long as the distribution of $[X_{jk}^h]$ over the population is observed accurately enough, the MNL model can still be estimated consistently without observing the choices of individual choosers. Modelers working in Wilson's tradition have been aware of the importance of the aggregate attribute information. Disaggregate demand modelers, on the other hand, have insisted on collecting information on $[\delta_j^h]$ without apparent knowledge that this information is necessary only to obtain the aggregate quantities $\mathring{N}_j$ and $\bar{X}_k$, which are the essential inputs in the estimation problem. It is, of course, true that in most cases the most efficient way to estimate $\mathring{N}_j$ and $\bar{X}_k$ is to observe $[\delta_j^h]$ and compute $\mathring{N}_j = \sum_h \delta_j^h$ and $\bar{X}_k = \sum_h \sum_j \delta_j^h X_{jk}^h$. But this need not always be the case, because aggregate measures such as travel cost, travel time, etc. may in many cases be estimated from observations on gasoline sales, transit station surveys, high-precision aerial photography techniques applied to traffic flow and from other sources. If a survey must be done, then the survey needed to obtain the attribute variation $[X_{jk}^h]$ and that needed to obtain the aggregates $\bar{X}_k$ and $\mathring{N}_j$ need not be the same one.

Corollary 2: The marginal utility of the kth generic attribute, β_k^*, is equal to the marginal change in the information level resulting from a marginal change in the observed aggregate value of the kth attribute over all alternatives and choosers choosing those alternatives.

Proof: Since $\beta_k^* = \partial \hat{U}_i^h / \partial X_{ik}^h$ (any i) and since $\lambda_k^* = \partial(-\mathscr{E})/\partial \bar{X}_k$, it follows from $\bar{\beta}^* = \bar{\lambda}^*$ that $\partial \hat{U}_i^h / \partial X_{ik}^h = \partial(-\mathscr{E})/\partial \bar{X}_k$ for each i. QED.

Corollary 3: The jth alternative-specific constant of the utility function β_{oj}^* is equal to the marginal change in the information level λ_{oj}^* resulting from a marginal change in the observed number of choosers choosing the jth alternative.

Proof: Since $\lambda_{oj}^* = \partial(-\mathscr{E})/\partial \mathring{N}_j$, it follows from $\bar{\beta}^* = \bar{\lambda}^*$ that $\beta_{oj}^* = \partial(-\mathscr{E})/\partial \mathring{N}_j$. QED.

Corollary 4: An attribute's contribution to the MNL model can be measured equivalently either as $\mathscr{E}^* - \mathscr{E}_k^*$ or as $\log \mathscr{L}^* - \log \mathscr{L}_k^*$ where $\mathscr{E}^*$ and $\mathscr{L}^*$ are the values of entropy and likelihood of the model estimated with all K attributes included and $\mathscr{E}_k^*$ and $\mathscr{L}_k^*$ are the values of entropy and likelihood of the model estimated with all K but the kth attribute.

Proof: $\mathscr{E}^* - \mathscr{E}_k^*$ and $\log \mathscr{L}^* - \log \mathscr{L}_k^*$ are two monotonically related measures of the marginal improvement in the information and log-likelihood achieved by the inclusion of the kth attribute. QED.

4. DATA AGGREGATION

In Sections 1 and 2 we treated the derivation of the MNL model from the disaggregate viewpoint. It has been assumed that each chooser's choices are known and given by a matrix $[\delta_j^h]$. We also assumed that the attribute values are given as $[X_{jk}^h]$ for each chooser h, each alternative j and each attribute k.

While utility-maximizing models are traditionally (but not exclusively) estimated from disaggregate data, maximum-entropy models are generally estimated from aggregated data. This, of course, is unnecessary; it is perfectly reasonable and just as easy computationally to maximize entropy for a sample of M aggregation units as it is to maximize entropy for a sample of M individual choosers. This fact follows directly from Section 2, in which the maximum-entropy model was cast in disaggregate format.

If choices are observed as aggregates, then only the mean value of each attribute k for alternative j, $\bar{X}_{jk} \equiv (\sum_h \delta_j^h X_{jk}^h)/\mathring{N}_j$ is observed and the MNL model can be written as,

$$P_i = \frac{\exp\left\{\beta_{oi}^A + \sum_k \beta_k{}^A \bar{X}_{ik}\right\}}{\sum_j \exp\left\{\beta_{oj}^A + \sum_k \beta_k{}^A \bar{X}_{jk}\right\}} \tag{26}$$

where P_i is the predicted relative frequency or expected choice probability for alternative i. The aggregated model can be estimated by maximizing the log-likelihood function

$$\log \mathscr{L} = \sum_j \mathring{N}_j \log P_j(\bar{\beta}^A) + \text{constant} \tag{27}$$

where $\mathring{N}_j$ is the number of choosers choosing alternative j and $P_j(\bar{\beta}^A)$ is (26).

The minimum-information formulation can be stated as,

$$\underset{[P_j]}{\text{Minimize}} -\mathscr{E} = \sum_j P_j \log P_j \tag{28}$$

$$\sum_j P_j = 1 \tag{29}$$

$$P_j = \mathring{N}_j/H; \qquad j = 1 \ldots J \tag{30}$$

$$\sum_j P_j \bar{X}_{jk} = \bar{X}_k/H; \qquad k = 1 \ldots K. \tag{31}$$

where H is the total number of choosers. Using Lagrangian minimization we can prove once again that the solution has the same form as (26) with λ_{oj}^A the Lagrange multiplier of (30) and $\lambda_k{}^A$ the Lagrange multiplier of (31). Furthermore, the first-order equations needed to maximize (26) are identical to those needed to maximize (28) subject to (29)–(31). It thus follows that $\lambda_{oj}^A = \beta_{oj}^A$ for each j, and $\lambda_k{}^A = \beta_k{}^A$ for each k.

Because disaggregate attribute information is not observed, the coefficients estimated from aggregated data will differ from those estimated from the underlying disaggregate data ($\bar{\beta}^{A*} \neq \bar{\beta}^*$ and $\bar{\lambda}^{A*} \neq \bar{\lambda}^*$); but it will be true that $\bar{\beta}^{A*} = \bar{\lambda}^{A*}$ and $\bar{\beta}^* = \bar{\lambda}^*$. The differences $\bar{\beta}^{A*} - \bar{\beta} = \bar{\lambda}^{A*} - \bar{\lambda}^*$ is known as aggregation bias. Empirical studies (see Anas 1981) demonstrate that meaningful estimates can be obtained when aggregation units are reasonably small, thus minimizing the aggregation bias.

5. THE DOUBLY CONSTRAINED GRAVITY MODEL IS A LOGIT MODEL OF JOINT ORIGIN-DESTINATION CHOICE

As a final culmination of the synthesis of behavioral choice modeling and non-behavioral information modeling, we will discuss the derivation of the gravity model and multiattribute generalizations of it, via both paradigms. Both derivations will be presented for the disaggregate case. The information-minimizing derivation is presented first.

Suppose $i = 1 \ldots I$ is a set of trip origin locations and $j = 1 \ldots J$ is a set of trip destination locations. An example of a trip origin location may be the location of an individual job, while an example of a trip destination location may be a dwelling. Let P_{ij}^h be the probability that chooser h will choose origin i and destination j, i.e. will choose to work at job i and reside at dwelling j. We seek the model which minimizes information subject to macro constraints.

$$\underset{\{P_{ij}^h\}}{\text{Minimize}} - \mathscr{E} = \sum_h \sum_i \sum_j P_{ij}^h \log P_{ij}^h \tag{32}$$

subject to:

$$\sum_i \sum_j P_{ij}^h = 1; \qquad h = 1 \ldots H \tag{33}$$

$$\sum_h \sum_j P_{ij}^h = \sum_h \sum_j \delta_{ij}^h; \qquad i = 1 \ldots I \tag{34}$$

$$\sum_h \sum_i P_{ij}^h = \sum_h \sum_i \delta_{ij}^h; \qquad i = 1 \ldots J \tag{35}$$

$$\sum_h \sum_i \sum_j P_{ij}^h X_{ijk}^h = \sum_h \sum_i \sum_j \delta_{ij}^h X_{ijk}^h; \qquad k = 1 \ldots K. \tag{36}$$

Letting λ_{oi}, $i = 1 \ldots I$, λ_{dj}, $j = 1 \ldots J$, and λ_k, $k = 1 \ldots K$ be the Lagrangian multipliers of (34)–(36), respectively, we derive the model as

$$P_{ij}^h = \frac{\exp\left(\lambda_{oi} + \lambda_{dj} + \sum_k \lambda_k X_{ijk}^h\right)}{\sum_m \sum_n \exp\left(\lambda_{om} + \lambda_{dn} + \sum_k \lambda_k X_{mnk}^h\right)} \tag{37}$$

where the λ's must be chosen to solve (34)–(36).

The behavioral derivation of the same model now follows. Suppose that the stochastic utility function has the form,

$$\hat{U}_{ij}^h = \beta_{oi} + \beta_{dj} + \sum_k \beta_k X_{ijk}^h + \epsilon_{ij}^h \tag{38}$$

and is the utility of choosing origin-destination pair (i, j) with ϵ_{ij}^h the random utility, β_{oi} the origin-specific unspecified utility, β_{dj} the destination-specific unspecified utility, and β_k the generic utility coefficients. If we assume that,

$$\text{Prob}\,(\epsilon_{ij}^h \leqq \epsilon) = \exp\left(-\exp\left[-\left(\frac{\pi^2}{6\sigma^2}\right)^{1/2} \epsilon\right]\right) \tag{39}$$

with mode zero and variance σ^2 for each (i, j), then the demand model can be derived from,

$$P_{ij}^h = \text{Prob.}\,[\hat{U}_{ij}^h > \hat{U}_{mn}^h; \qquad \forall (m, n) \neq (i, j)] \tag{40}$$

and has the form,

$$P_{ij}^h = \frac{\exp\left(\beta_{oi} + \beta_{dj} + \sum_k \beta_k X_{ijk}^h\right)}{\sum_m \sum_n \exp\left(\beta_{om} + \beta_{dn} + \sum_k \beta_k X_{mnk}^h\right)}. \tag{41}$$

Once again, (41) and (37) are identical, and estimation will yield $\bar{\beta} = \bar{\lambda}$. In this formulation there are $I \cdot J$ alternatives and $\beta_{oi} + \beta_{dj} = \lambda_{oi} + \lambda_{dj}$ is the alternative-specific constant for the (i, j)th alternative.

Suppose now that an origin, i, is not a distinct job (or other) alternative, but a zone (spatial aggregation) of alternatives; and a destination, j, is likewise an aggregation of home (or other) alternatives. Let there by $i = 1 \ldots I$ zones of origin and $j = 1 \ldots J$ zones of destination. If spatial alternatives are so aggregated, then the attribute measures $\bar{X}_{ijk}$ are the means for attribute k over the choosers choosing any specific origin within zone i and any specific destination within zone j, namely $\bar{X}_{ijk} = \sum_h \delta_{ij}^h X_{ijk}^h / \sum_h \delta_{ij}^h$. The model of joint origin-destination choice now becomes a model of joint origin-destination *zone pair* choice, and must be written in aggregated form, as

$$P_{ij} = \frac{\exp\left\{\lambda_{oi}^A + \lambda_{dj}^A + \sum_k \lambda_k^A \bar{X}_{ijk}\right\}}{\sum_m \sum_n \exp\left\{\lambda_{om}^A + \lambda_{dn}^A + \sum_k \lambda_k^A \bar{X}_{mnk}\right\}} \tag{42}$$

where λ_{oi}^A, λ_{dj}^A and λ_k^A are the coefficients to be estimated with aggregation error. To derive (42) we must pose the following information-minimization problem, which seeks to find $[P_{ij}]$, the matrix of relative frequencies or expected choice probabilities. The formulation is,

$$\underset{[P_{ij}]}{\text{Minimize}} - \mathscr{E} = \sum_i \sum_j P_{ij} \log P_{ij} \tag{43}$$

$$\sum_j P_{ij} = O_i/H; \qquad i = 1 \ldots I \tag{44}$$

$$\sum_i P_{ij} = D_j/H; \qquad j = 1 \ldots J \tag{45}$$

$$\sum_i \sum_j P_{ij} \bar{X}_{ijk} = \bar{X}_k/H; \qquad k = 1 \ldots K \tag{46}$$

with λ_{oi}^A, λ_{dj}^A and λ_k^A the Lagrange multipliers of (44)–(46), respectively. Equation (42) will satisfy constraints (44)–(46) if the Lagrange multipliers are appropriately defined. It must be true, of course, that $\sum_i O_i = \sum_j D_j = H$, where O_i is the number of trips (choices) originating at i and D_j the number of trips (choices) terminating at j.

The relationship between the MNL model (42) also derived from (43) subject to (44)–(46) and the conventional gravity model follows if we first recall that $N_{ij} = HP_{ij}$—i.e., N_{ij} is the expected number of choosers choosing (i, j), and thus the expected number of trips (or exchanges) between i and j. In the workplace/residence choice example, N_{ij} is the expected number of commutes between i and j. Next, we make the definitions,

$$\exp(\lambda_{oi}^A) \equiv O_i \Big/ \sum_j \exp\left(\lambda_{dj}^A + \sum_k \lambda_k^A \bar{X}_{ijk}\right), \tag{47}$$

$$\exp(\lambda_{dj}^A) \equiv D_j \Big/ \sum_i \exp\left(\lambda_{oj}^A + \sum_k \lambda_k^A \bar{X}_{ijk}\right) \tag{48}$$

and

$$A_i \equiv \exp(\lambda_{oi}^A)/O_i \tag{49}$$

$$B_j \equiv \exp(\lambda_{dj}^A)/D_j. \tag{50}$$

From these definitions we can write,

$$N_{ij} = (HP_{ij}) = A_i B_j O_i D_j \exp\left(\sum_k \lambda_k^A \bar{X}_{ijk}\right) \tag{51}$$

with

$$A_i = \left[\sum_j B_j D_j \exp\left(\sum_k \lambda_k^A \bar{X}_{ijk}\right)\right]^{-1} \tag{52}$$

$$B_j = \left[\sum_i A_i O_i \exp\left(\sum_k \lambda_k^A \bar{X}_{ijk}\right)\right]^{-1}. \tag{53}$$

This model (51)–(53) is precisely Wilson's (1967) doubly-constrained entropy (or gravity) model with several attributes in $\exp(\cdot)$. It is identical to eqn (1) if we simply define $G_{ij} = A_i B_j$ and replace $\exp(\cdot)$ with a generalized function of distance, $f(d_{ij})$. As a historical curiosity, the Newtonian gravity model is obtained directly from (51) by defining $G_{ij} = A_i B_j$ *and* letting $\exp(\cdot) = \lambda^A \log d_{ij}$ where $\lambda^A = -2$ and d_{ij} is the average distance between i and j, the only attribute in the "utility function". The result is,

$$N_{ij} = G_{ij} \frac{O_i D_j}{d_{ij}^2}. \tag{54}$$

We have thus come full circle and shown that the doubly-constrained gravity model (51) *is* a multinomial logit model of joint origin-destination choice consistent with stochastic utility maximization up to some aggregation error in the estimated coefficients. Furthermore, we have shown that the same model without any aggregation error is derivable in disaggregate form (41) and can be thus estimated. To estimate the gravity model correctly one must, in fact, estimate a multinomial logit model of joint origin-destination choice from disaggregate data.

6. BEHAVIORAL VERSUS INFORMATIONAL APPROACHES TO MODEL SPECIFICATION

We have proved that stochastic utility maximization based on assumptions A1–A3 yields results which are identical to information minimization based on assumptions A4 and A5. The former approach is "behavioral" while the latter is purely informational, but the result is the same. This means that analysts who are following Wilson's paradigm and others following McFadden's are engaged in precisely the same endeavor, of which they have different—indeed, opposed—views. If these two analysts are given the same data (disaggregate or aggregate) and asked to produce the same model specification, then both will estimate the same model. If, on the other hand, each is allowed to seek and find the "best" specification, then each may arrive at a different answer. The reason for this is that the model-specification criteria employed by the two analysts will probably differ. The behavioral analyst is very likely to begin the work with strong preconceptions of what goes into a utility function and to select only those attributes, leaving out superfluous attributes and unsatisfactory proxy attributes. The pure information theorist is an "agnostic" in comparison to the behavioral "believer". Information theory *per se* does not contain any insights as to what *should* go into a model. Nevertheless, the context of a particular problem does in most cases provide clues as to which attributes are "explanatory" and which are not, and also as to the correct sign a particular attribute coefficient should obtain.

Information theory is at once more general than utility maximization: since the MNL model is consistent with both utility maximization and information theory, the information theorist will in many instances succeed in producing models which are acceptable to the economist who is operating from a behavioral viewpoint. The choice of a particular model is ultimately conditioned by *prior belief* (value judgments) about what attributes to enter into the utility function or what constraints to impose in the information-minimizing approach.

Historically, the different uses of data within the two paradigms has acted as a communication barrier, obscuring the full equivalence of the two methods. Wilson's pioneering contribution took aggregative gravity models as the point of departure and did not sufficiently emphasize the applicability of information theory at the disaggregate level. Had this been done, an early synthesis with behavioral modeling could have been achieved. The behavioral modelers took a suspicious and misguided view of entropy maximizing modeling, mistaking what is merely a problem in aggregation error for a more serious difference in mathematical structure.

7. CONCLUSIONS

This paper shows that it is no longer reasonable or excusable to claim that entropy and gravity models are *inherently* less "behavioral" than stochastic utility models of discrete choice

and multinomial logit in particular. The two approaches are two equivalent views of the same problem. The fact that models estimated within one approach have yielded results quite different from similar models estimated within the other approach is not due to any inherent difference between the two approaches, but is due *entirely* to differences in the use of data and its aggregation and in differences in value judgments used in specifying the explanatory attributes. These differences are determined entirely by historical intertia and intellectual bias on the part of investigators.

It is important to recall that, although information theory and discrete choice theory overlap in the MNL model, *each* of the two approaches is more general than it appears from this context. Information minimization is a powerful principle which does not subscribe to any behavioral postulates. It is thus applicable to a wide range of problems as a tool for generating most probable predictions subject to available information and in the absence of any behavioral structure. Discrete choice theory, on the other hand, is a powerful tool for exploring the validity of a potentially large number of specific behavioral postulates about preferences and their stochastic distribution. The MNL model corresponds to one subset of such postulates; multinomial probit (MNP), generalized extreme value (GEV) and other choice models correspond to other postulates.

It would be naive to think that the equivalence between information minimization and discrete choice theory ends with the MNL model. The MNL model is derived by minimizing information subject to *linear* macro-constraints. Nonlinear constraints reflecting higher moment properties of the distribution of certain attributes will result in models which are more complex but possibly also more realistic. One cannot help but ask: what additional macro-behavioral constraints must be imposed on information minimization in order to derive the multinomial probit model?

REFERENCES

Anas A. (1975) Empirical calibration and testing of a simulation model of residential location, *Envnmt Plan. A.* 7, 899–920.

Anas A. (1981) The estimation of multinomial logit models of joint location and mode choice from aggregated data, *J. Reg. Sci.* **21** 223–242.

Ben-Akiva M. E. (1973) Structure of passenger travel demand models. Ph.D. dissertation. Department of Civil Engineering. MIT, Cambridge, Mass.

Los M. (1979) Discrete choice modeling and disequilibrium in land use and transportation planning, *Working Paper No.* 137. Centre de recherche sur les transports, Universitè de Montreal, August.

McFadden D. (1973) Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics* (Edited by P. Zarembka). Academic Press, New York.

Warner S. L. (1962) *Stochastic Choice of Mode in Urban Travel: A Study in Binary Choice.* Northwestern University Press, Evanston, Illinois.

Williams, H. C. W. L. (1977) On the formation of travel demand models and economic evaluation measures of user benefit. *Envnmt Plan. A.* **9**, 285–344.

Wilson A. G. (1967), A statistical theory of spatial distribution models, *Transpn Res.* **1**, 253–269.