Replication/prediction problems in the Journey-to-work

Kurt Jörnsten[1], Inge Thorsen[2], and Jan Ubøe[1]

[1] Norwegian School of Economics and Business Administration
Helleveien 30, N-5045 Bergen, Norway.

[2] Stord/Haugesund College, Bjørnsonsgate 45, 5528 Haugesund, Norway.

ABSTRACT. In this paper we will consider gravity models for journeys-to-work. In applications of the theory it is sometimes assumed that the parameters in such models are fixed. We will provide examples to show that this is not always a reasonable assumption, for instance when the model is applied to predict how changes in the road transportation network influence commuting flows. Models where the parameters are subject to change usually comply with C-efficiency and random utility theory.

1. Introduction

Economic evaluations of investments in transportation infrastructure in general call for predictions on how traffic flows are affected. Most commonly such predictions are based on models belonging to the gravity modeling tradition. Consider the following doubly-constrained version of a gravity model:

$$T_{ij}^G = A_i B_j e^{-\beta d_{ij}} \qquad i,j = 1,\ldots,N \tag{1.1}$$

$$\sum_{k=1}^{N} T_{ik}^G = L_i \qquad \sum_{k=1}^{N} T_{kj}^G = E_j \qquad i,j = 1,\ldots,N \tag{1.2}$$

$$\sum_{i=1}^{N} L_i = \sum_{j=1}^{N} E_j \tag{1.3}$$

Here T_{ij}^G denotes the number of travelers from origin i to destination j, see Section 2 for definitions of the various other symbols. This doubly-constrained model formulation is constructed for trip distribution problems. For a discussion on the theoretical foundation of this model, see for instance Erlander and Stewart (1990) and/or Sen and Smith (1995).

As a first step the model is normally used to calibrate the parameter β from observations representing the current state of the system. This parameter is traditionally interpreted to reflect how individuals in general respond to distance in the relevant geography, and the model offers an explanation to the observed traffic flow pattern. In many applications of the model, however, the problem is to predict how the traffic flow pattern responds to a specific exogenous change in the system. The prediction of a new state is then based on the assumption that the parameter β is autonomous to the exogenous change. This paper primarily deals with the validity of this assumption, and with the interpretation of the distance deterrence parameter and the interaction model in general.

Traditionally the distance deterrence parameter in gravity models was interpreted as a behavioral measure. In the late seventies and early eighties this interpretation was challenged both by theoretical analysis and by empirical findings. Based on airline passenger interaction between the hundred largest cities in the US, Fotheringham (1981) offered origin-specific estimates of the distance deterrence parameter. Within an unconstrained model formulation Fotheringham (1981) found considerable spatial variation in parameter estimates, and he found that this variation depended systematically on the accessibility of an origin within the configuration of cities. Those findings initiated a debate focusing primarily on the impact of spatial structure characteristics in applications of gravity models. Fotheringham (1983a) found that the spatial variation in parameter estimates were considerably reduced in a production constrained modeling framework, which corresponds to a model formulation with an exogenuously given number of trips originating from each zone ($\sum_{k=1}^{N} T_{ik}^{G} = L_i$). This introduces a balancing factor that to some degree implicitly captures the possibility that interaction flows are influenced by the accessibility pattern of the alternative destinations. Several succeeding studies have demonstrated that model performance is improved if the accessibility of the destinations is explicitly taken into account, see for instance Fotheringham (1983b, 1984, 1986), Ishikawa (1987), Desta and Pigozzi (1991) and Thorsen and Gitlesen (1998). The introduction of destination accessibility defines the so called competing destinations model as a specific variant within the family of gravity interaction models.

The discussion concerning destination accessibility was motivated by the general idea that the traditional gravity model is misspecified, since it ignores relevant characteristics of the spatial structure. If strong agglomeration or competition effects are present, then the distribution of trips will be affected by the clustering system of destinations in addition to distance. The effect of omitted spatial structure variables is that estimates of the distance deterrence parameter are biased, and not autonomous to for instance variations in space or changes in the transportation network. Some studies focus on other aspects of the spatial structure than the position of a potential destination relative to the other destination alternatives. Baxter (1983) derives rather complex formulas for specification errors in a traditional

spatial interaction model, and finds that the bias in parameter estimates in general depends on many characteristics of spatial structure. Fik and Mulligan (1990) and Fik et al. (1992) extend the relevant family of spatial interaction models beyond the pure competing destinations model. They find that taking special account of the hierarchical order of potential destinations and to the number of intervening opportunities adds significantly to the explanatory power.

Some contributions to the literature are critical to the competing destinations approach, see for instance Ewing (1986). Gordon (1983) claims that the competing destinations model "differ only in trivial and arbitrary particulars from the conventional doubly-constrained gravity model" and that "spatial variation in distance deterrence should focus on functional and economic differences between areas rather than simply on map pattern or physical accessibility". In more recent studies of spatial interaction modeling focus has in fact tended to change towards a discussion of economic and behavioral aspects. Fotheringham (1986) derives the competing destinations model from random utility theory by including destination accessibility as an additive separable component of the utility function. If shopping is the relevant kind of spatial interaction destination accessibility enters as an attribute in a utility context, since it represents the availability to alternative opportunities in destinations nearby. Accessible and spatially complementary destinations might be most attractive as a consequence of the economies of scale in the information flow and the underlying trip structure. This corresponds to the effect of economic substitution between goods and services in alternative destinations, see Lo (1991). Other approaches stress the importance of distinguishing between the universal and the true choice set of individual decisionmakers. Applied to shopping problems Fotheringham (1988) interprets the competing destinations model to result from a two-stage decisionmaking process, corresponding to the ideas of probabilistic choice set generation in Manski (1977). First, the decisionmakers select the set of alternatives which are relevant destination choices. Second, a specific destination is selected from this set of alternatives. Accessibility affects the hierarchical structure, while variables which appear through the structural part of the utility function affect the second stage of the decisionmaking process. Thill (1992) questions the behavioral base of Fotheringham's (1988) approach, arguing that "the behavioral soundness of the model would require other factors to be considered, if only the distance from home". Pellegrini et al. (1997) demonstrate that a misspecification of choice sets might produce misleading parameter estimates and predictions; parameter estimates vary systematically with respect to the definition of choice sets in shopping destination models. Based on numerical experiments Thill and Horowitz (1997) found that substantial prediction errors might result if the presence of limited time budgets is ignored in the model specification.

It is well known in the literature that the family of gravity models can be justified from random utility maximization. Such models can also be derived from a principle of C-efficiency with respect to a linear cost function $C = \sum_{i,j} c \cdot d_{ij}$, where c

is a constant traveling cost per unit of distance. Hence, the models are based on a sound foundation, and we think that the referred research has generated model formulations with satisfying explanatory power to a specific pattern of spatial interaction flows. As indicated above, however, we are more concerned by the practice of applying such models for prediction purposes. Even if spatial structure is correctly specified we question the idea of a constant distance deterrence parameter, autonomous to the relevant exogenous changes in the system. Our concern is based on arguments that have not been addressed in the referred literature. C-efficiency and random utility maximization deal with systems that are static in the sense that distances are fixed. In this paper, however, we will see that changes in the system might affect individual preferences and behavior. In replication/prediction problems distances are subject to change, and a relevant model should be equipped with hypotheses on how the behavior is affected by such changes. To keep analysis as simple as possible, our points are made clear within the framework of a conventional gravity model ((1.1), (1.2) and (1.3)), with no attempts to account for the possibility that additional measures of the spatial structure contribute to explain the distribution of trips. In this paper we also restrict our analysis to consider commuting flows.

The paper is organized as follows: In Section 2 we recall the basic properties of C-efficiency with reference to Erlander and Smith (1990). In particular we prove that given a linear cost $C = \sum_{i,j} c d_{ij}$, an activity matrix A and a probability measure $\mathbf{q} > 0$, there is an infinite number of trip distribution models that are C-efficient w.r.t. A. In particular the β-parameter in a standard gravity model need not be unique, given C, A and $\mathbf{q}$. In Section 3 we consider replication/prediction problems, and prove that any non-degenerate trip distribution model comply with C-efficiency/random utility if the cost structure is chosen in an appropriate way. Hence C-efficiency/random utility is not sufficient to the determine the outcome of a replication/prediction problem. To determine a unique outcome it is necessary to equip the system with an additional behavioral hypothesis. Constancy of the β-parameter is one such additional assumption, but it is not the only one. In Section 4 we consider examples where an assumption of a constant β-parameter may lead to completely wrong predictions. In Section 5 we briefly discuss some alternative behavioral assumptions and compare the outcomes in the different cases. Finally in Section 6 we offer some concluding remarks.

2. C-efficiency

In this section we will briefly recall the main properties of C-efficiency, see Erlander and Smith (1990). To simplify the discussion, we will consider a slightly stronger version where activities need not be integer valued. This keeps all the essential features of the original concept, while at the same time admitting a more transparent exposition. The basic core of the argument is nevertheless the same as in Erlander and Smith (1990), so we cannot claim originality on that part.

Consider a choice between M alternatives $a_1, \dots, a_M$, let $\mathbf{z} = (z_1, \dots, z_M)$ denote observed frequencies of these alternatives, and let $n = \sum_{k=1}^{M} z_i$ denote the total number of choices that are made. We will assume that choices are made from a probability distribution $\mathbf{P} = (P_1, \dots, P_M)$. If the choices are independent and n is very large, it is reasonable to assume that $\mathbf{z} \approx \mathbf{P} \cdot n$. The alternatives are equipped with costs $\mathbf{c} = (c_1, \dots, c_M)$, and hence $\mathbf{z}$ gives rise to a total cost $C = \sum_{k=1}^{M} c_k z_k$.

EXAMPLE 2.1

In a trip distribution model in a system with N nodes, the alternatives are trips with origin $i = 1, \dots, N$ to destination $j = 1, \dots, N$; a total of $M = N^2$ alternatives. To simplify notation, we introduce the following identification:

If $\mathbf{B} = \{B_{ij}\}_{i,j=1}^{N}$ is an $N \times N$ matrix, we let

$$\vec{\mathbf{B}} = (B_{11}, \dots, B_{1N}, B_{21}, \dots, B_{2N}, \dots, B_{NN}) \in \mathbb{R}^{N^2} \tag{2.1}$$

Hence if we associate a cost $c_{ij} = c \cdot d_{ij}$ from traveling the distance d_{ij} between origin i and destination j, we can define a cost matrix $\mathbf{c} = \{c_{ij}\}_{i,j=1}^{N}$, which in turn defines a cost vector $\vec{\mathbf{c}}$ through (2.1). Note that with these conventions, the total cost C can be expressed on the form

$$C = \sum_{i,j=1}^{N} c_{ij} z_{ij} = \vec{\mathbf{c}} \cdot \vec{\mathbf{z}} \tag{2.2}$$

In applications of this theory, we will only consider frequencies $\mathbf{z} = \{z_{ij}\}_{i,j=1}^{N}$ that are consistent with the marginal restrictions on the system, i.e.

$$\sum_{i=1}^{N} z_{ij} = E_j \qquad \sum_{j=1}^{N} z_{ij} = L_i \qquad i, j = 1, \dots, N \tag{2.3}$$

where E_j denotes the number of employment opportunities in destination j, and L_i denotes the number of workers residing in origin i. For consistency we always assume that $\sum_{i=1}^{N} L_i = \sum_{j=1}^{N} E_j$. To write the restrictions in (2.3) on a form that is more suitable for analysis, we consider a $2N \times N^2$ matrix $\mathbf{A}$, defined as follows:

Case 1: $1 \le k \le N$

$$A_{kl} = \begin{cases} 1 & \text{if } l = k + iN, \text{ where } i = 0, \dots, N-1 \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

Case 2: $N+1 \leq k \leq 2N$

$$A_{kl} = \begin{cases} 1 & \text{if } (k-N-1)N < l \leq (k-N)N \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

If we define a vector $\mathbf{b} \in \mathbb{R}^{2N}$ by

$$\mathbf{b} = (E_1, \ldots, E_N, L_1, \ldots, L_N) \tag{2.6}$$

then one can verify that (2.3) is equivalent to the matrix equation

$$\mathbf{A}\vec{\mathbf{z}}^{\perp} = \mathbf{b}^{\perp} \tag{2.7}$$

where $\vec{\mathbf{z}}^{\perp}$ denotes the transpose of the vector $\vec{\mathbf{z}}$, i.e.

$$\vec{\mathbf{z}}^{\perp} = \begin{bmatrix} \vec{z}_1 \\ \vec{z}_2 \\ \vdots \\ \vec{z}_{N^2} \end{bmatrix} \tag{2.8}$$

EXAMPLE 2.2
Here is how (2.3)/(2.7) looks like for a 2×2 system:

$$\mathbf{T} = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} z_{11} \\ z_{12} \\ z_{21} \\ z_{22} \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \\ L_1 \\ L_2 \end{bmatrix} \tag{2.9}$$

We then consider the following definition: We say that two frequency vectors $\vec{u}, \vec{v}$ are activity equivalent under $\mathbf{A}$ if

$$\mathbf{A}\vec{u} = \mathbf{A}\vec{v} \tag{2.10}$$

In the context of trip distribution models, with $\mathbf{A}$ defined by (2.4-5), activity equivalence is then just to say that both frequencies satisfy the standard marginal restrictions in (2.3).

Let $\mathbf{P} = (P_1, \ldots, P_M)$ be any probability measure on the set of alternatives, and let $C = \vec{\mathbf{c}} \cdot \vec{\mathbf{z}}$ denote the total cost associated with the frequency $\mathbf{z}$. Loosely speaking, the

efficiency principle defines that states with small costs should be more probable, i.e., if

$$\vec{\mathbf{c}} \cdot \vec{u} \leq \vec{\mathbf{c}} \cdot \vec{\nu} \qquad \Rightarrow \qquad \prod_{k=1}^{M} P_k^{u_k} \geq \prod_{k=1}^{M} P_k^{\nu_k} \tag{2.11}$$

or equivalently

$$\vec{\mathbf{c}} \cdot \vec{u} \leq \vec{\mathbf{c}} \cdot \vec{\nu} \qquad \Rightarrow \qquad \sum_{k=1}^{M} \ln[P_k] \cdot u_k \geq \sum_{k=1}^{M} \ln[P_k] \cdot \nu_k \tag{2.12}$$

Strictly speaking, this only makes sense when u and ν are frequencies, i.e., non-negative integers. To avoid some technical complications, we exploit the following definition: A probability measure $\mathbf{P}$ is called strongly C-efficient under $\mathbf{A}$ if for all probability measures $\mathbf{p} = (p_1, \ldots, p_M)$ and $\mathbf{q} = (q_1, \ldots, q_M)$

$$\vec{\mathbf{c}} \cdot \mathbf{p} \leq \vec{\mathbf{c}} \cdot \mathbf{q},\ \mathbf{A}\mathbf{p}^{\perp} = \mathbf{A}\mathbf{q}^{\perp} \qquad \Rightarrow \qquad \sum_{k=1}^{M} \ln[P_k] \cdot p_k \geq \sum_{k=1}^{M} \ln[P_k] \cdot q_k \tag{2.13}$$

If we define $\ln[\mathbf{P}] = (\ln[P_1], \ldots, \ln[P_M])$, we can write this on the form

$$\vec{\mathbf{c}} \cdot \mathbf{p} \leq \vec{\mathbf{c}} \cdot \mathbf{q},\ \mathbf{A}\mathbf{p}^{\perp} = \mathbf{A}\mathbf{q}^{\perp} \qquad \Rightarrow \qquad \ln[\mathbf{P}] \cdot \mathbf{p} \geq \ln[\mathbf{P}] \cdot \mathbf{q} \tag{2.14}$$

Let $\sigma = (1, 1, \ldots, 1) \in \mathbb{R}^M$. Then $\mathbf{x} = (x_1, \ldots, x_M) \geq 0$ is a probability measure if and only if $\sigma \cdot \mathbf{x} = 1$. Choose any probability measure $\mathbf{q} > 0$, and consider the LP-problem

$$\begin{aligned} &\min_{\mathbf{x}}\ \ln[\mathbf{P}] \cdot \mathbf{x} \\ &\sigma \cdot \mathbf{x} = \sigma \cdot \mathbf{q} \\ &\mathbf{A}\mathbf{x}^{\perp} = \mathbf{A}\mathbf{q}^{\perp} \\ &\vec{\mathbf{c}} \cdot \mathbf{x} \leq \vec{\mathbf{c}} \cdot \mathbf{q} \\ &\mathbf{x} \geq 0 \end{aligned} \tag{2.15}$$

If $\mathbf{P}$ is strongly C-efficient under $\mathbf{A}$, it is trivial to see that (2.15) must have the solution $\mathbf{x}^* = \mathbf{q} > 0$ (otherwise the pair $(\mathbf{x}^*, \mathbf{q})$ would violate (2.14)). Obviously, (2.15) has a dual problem which can be expressed as follows: $\mathbf{y} = (y_0, y_1, \ldots, y_{2N}, y_{2N+1})$

$$\begin{aligned} &\max_{\mathbf{y}}\ \mathbf{q}\,(\sigma^{\perp}, \mathbf{A}^{\perp}, -\vec{c}^{\perp})\mathbf{y}^{\perp} \\ &(\sigma^{\perp}, \mathbf{A}^{\perp}, -\vec{c}^{\perp})\mathbf{y}^{\perp} \leq \ln[\mathbf{P}]^{\perp} \\ &y_k \in \mathbb{R}, k = 0, 1, \ldots, 2N \\ &y_{2N+1} \in \mathbb{R}^{+} \end{aligned} \tag{2.16}$$

Here we have used the shorthand notation

$$(\sigma^{\perp}, \mathbf{A}^{\perp}, -\vec{c}^{\perp}) = \begin{bmatrix} 1 & a_{11} & a_{21} & \cdots & a_{2N,1} & -c_1 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 1 & a_{1,N^2} & a_{2,N^2} & \cdots & a_{2N,N^2} & -c_{N^2} \end{bmatrix} \tag{2.17}$$

Since $\mathbf{x}^* > 0$, all slack variables in the dual problem must be zero. Hence

$$\ln[\mathbf{P}]^{\perp} = y_0\sigma^{\perp} + \mathbf{A}^{\perp}(y_1, \ldots, y_{2N})^{\perp} - y_{2N+1}\vec{\mathbf{c}}^{\perp} \tag{2.18}$$

Conversely, if

$$\ln[\mathbf{P}]^{\perp} = y_0\sigma^{\perp} + \mathbf{A}^{\perp}(y_1, \ldots, y_{2N})^{\perp} - y_{2N+1}\vec{\mathbf{c}}^{\perp} \tag{2.19}$$

then for all pairs $(\mathbf{p}, \mathbf{q})$ s.t. $\vec{\mathbf{c}} \cdot \mathbf{p} \leq \vec{\mathbf{c}} \cdot \mathbf{q}$, $\mathbf{A}\mathbf{p}^{\perp} = \mathbf{A}\mathbf{q}^{\perp}$, we have (using $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}\mathbf{v}^{\perp}$)

$$\begin{aligned} &\ln[\mathbf{P}] \cdot \mathbf{p} - \ln[\mathbf{P}] \cdot \mathbf{q} \\ &= (y_0\sigma^{\perp} + \mathbf{A}^{\perp}(y_1, \ldots, y_{2N})^{\perp} - y_{2N+1}\vec{\mathbf{c}}^{\perp})^{\perp}(\mathbf{p}^{\perp} - \mathbf{q}^{\perp}) \\ &= y_0(\sigma \cdot \mathbf{p} - \sigma \cdot \mathbf{q}) + (y_1, \ldots, y_{2N})(\mathbf{A}\mathbf{p}^{\perp} - \mathbf{A}\mathbf{q}^{\perp}) + y_{2N+1}(\vec{\mathbf{c}} \cdot \mathbf{q} - \vec{\mathbf{c}} \cdot \mathbf{p}) \\ &= y_{2N+1}(\vec{\mathbf{c}} \cdot \mathbf{q} - \vec{\mathbf{c}} \cdot \mathbf{p}) \geq 0 \end{aligned} \tag{2.20}$$

Hence $\ln[\mathbf{P}] \cdot \mathbf{p} \geq \ln[\mathbf{P}] \cdot \mathbf{q}$, proving that $\mathbf{P}$ is strongly C-efficient under $\mathbf{A}$.

We have proved the following: If $\mathbf{P}$ is a probability measure that is strongly C-efficient under $\mathbf{A}$, then

$$\ln[\mathbf{P}]^{\perp} = y_0\sigma^{\perp} + \mathbf{A}^{\perp}(y_1, \ldots, y_{2N})^{\perp} - y_{2N+1}\vec{\mathbf{c}}^{\perp} \tag{2.21}$$

What is not clear, however, is the following: Given $C, \mathbf{A}$ and $\mathbf{q} > 0$, is it then always possible to find a probability measure $\mathbf{P}$ s.t.

i) $\mathbf{P}$ is strongly C-efficient under $\mathbf{A}$
ii) $\mathbf{A}\mathbf{P}^{\perp} = \mathbf{A}\mathbf{q}^{\perp}$ (corresponding to (2.7))

The difficulty is now if we can find $\mathbf{y} = (y_0, y_1, \ldots, y_{2N}, y_{2N+1})$ with $y_{2N+1} > 0$, s.t. $\mathbf{P}$ defined by (2.21) is a probability measure satisfying $\mathbf{A}\mathbf{P}^{\perp} = \mathbf{A}\mathbf{q}^{\perp}$. Choose and fix any $y_{2N+1} > 0, y_0 \in \mathbb{R}$. Define $\beta = cy_{2N+1}$, and consider a gravity model

$$P_{ij} = R_i S_j e^{-\beta d_{ij}} \tag{2.22}$$

with the side conditions

$$\sum_{i=1}^{N} P_{ij} = \sum_{i=1}^{N} q_{ij} \qquad \sum_{j=1}^{N} P_{ij} = \sum_{j=1}^{N} q_{ij} \tag{2.23}$$

Choose any pair R, S solving these equations, and define

$$y_k = \begin{cases} \ln[S_k] - y_0 & \text{if } 1 \le k \le N \\ \ln[R_k] & \text{if } N+1 \le k \le 2N \end{cases} \tag{2.24}$$

It is then tedious but straightforward to verify that with these definitions we get a probability measure satisfying $\mathbf{AP}^{\perp} = \mathbf{Aq}^{\perp}$.

Erlander and Smith (1990) observe that given $\mathbf{P}$, then $y_0, \ldots, y_{2N+1}$ are (usually) unique. From the argumentation above, we see that in the converse problem, i.e., given $C, \mathbf{A}$ and $\mathbf{q} > 0$, the probability $\mathbf{P}$ is not unique. In particular every different choice of y_{2N+1} corresponds to a different choice of β in the gravity model. When distances within a network are subject to change, there is hence little reason why the values on the β-parameter could not change as well.

3. Predictions

We now consider a given network with a distance matrix $\mathbf{d}^{\text{original}}$. We wish to consider a change in the network giving rise to a new distance matrix $\mathbf{d}^{\text{new}}$. Consider any given pair of positive real numbers β_1, β_2, and let $\beta : \mathbb{R}^N \to \mathbb{R}$ be any continuous function such that

$$\beta[\mathbf{d}^{\text{original}}] = \beta_1 \qquad \beta[\mathbf{d}^{\text{new}}] = \beta_2 \tag{3.1}$$

Choose any probability measure $q > 0$ and consider a gravity model on the form

$$P_{ij} = R_i S_j e^{-\beta[\mathbf{d}] d_{ij}} \tag{3.2}$$

with the side conditions

$$\sum_{i=1}^{N} P_{ij} = \sum_{i=1}^{N} q_{ij} \qquad \sum_{j=1}^{N} P_{ij} = \sum_{j=1}^{N} q_{ij} \tag{3.3}$$

If the cost function is of the form $c_{ij} = c\, d_{ij}$, then clearly $\mathbf{P}$ defined from (3.1) and (3.2) is strongly C-efficient with respect to the marginal conditions (2.3). One hence

cannot necessarily conclude that β is a universal constant, even if the system varies continuously and is strongly C-efficient for all $\mathbf{d}$.

To proceed one step further, one can consider gravity models on the form

$$P_{ij} = R_i S_j e^{-f_{ij}[\mathbf{d}]} \tag{3.4}$$

Choose and fix any $\mathbf{d} \geq 0$. We can then think of $\tilde{d}_{ij} = f_{ij}[\mathbf{d}]$ as a generalized distance between origin i and destination j, and consider a generalized cost function of the form $c_{ij} = c\,\tilde{d}_{ij}$. The model in (3.4) is hence strongly C-efficient with respect to this generalized cost regardless of the choice of $\mathbf{d}$.

If $\mathbf{d} \mapsto \mathbf{P}[\mathbf{d}]$ denotes any non-degenerate trip distribution model that is consistent with respect to the marginal constraints (2.3), we may of course consider a gravity model on the form

$$P_{ij}[\mathbf{d}] = R_i S_j e^{-\ln\left[\frac{1}{P_{ij}[\mathbf{d}]}\right]} \tag{3.5}$$

In this sense we can always define generalized costs such that any non-degenerate trip distribution model is globally C-efficient w.r.t. to these costs.

The argumentation above applies to C-efficiency. A similar set of arguments applies to a random utility maximizing. With an appropriate choice of utility function, any non-degenerate trip distribution model complies with this principle as well. If we restrict attention to utilities defined in terms of a linear cost function, however, we end up with a standard gravity model. Whenever a change is introduced in the network, there is little reason to exclude the possibility that the parameter of the extreme value distribution could change as well. Hence one cannot argue that the β-parameter is necessarily constant in such systems.

4. Predictions made from constant parameters

Ubøe (2001) discusses several problems with the gravity model in aggregate systems. To examine this within the context of C-efficiency/random utility, we elaborate a bit further on the following example from Ubøe (2001):

EXAMPLE 4.1

Consider a region with two towns and two different population groups. In the first group we have 3000 workers and 5500 employment opportunities in town 1, while in town 2 we have 7000 workers and 4500 employment opportunities. In the second group there are 7000 workers and 4500 employment opportunities in town

1, while in town 2 there are 3000 workers and 5500 employment opportunities. The two population groups are non-interacting, i.e., the people in the first category cannot take work in the second category and vice versa. We will assume that both categories behave according to a standard gravity model as in (1.1-3), both with the same value on the deterrence parameter; $\beta = 0.03$.

Assume that initially the distance between the two towns is $d_{12} = 80$ (km). Using the gravity model on each category and adding the result together, we end up with an aggregated trip distribution matrix:

$$\mathbf{T}_{80} = \begin{bmatrix} 7415 & 2585 \\ 2585 & 7415 \end{bmatrix} \tag{4.1}$$

This aggregate system is strongly C-efficient w.r.t a linear cost function. The system has a perfect replication by a standard gravity model on the aggregated data, i.e.,

$$L_1 = 10000, L_2 = 10000, E_1 = 10000, E_2 = 10000 \tag{4.2}$$

and the replicating parameter is $\beta = 0.0137$. We now change the distance between the towns to $d_{12} = 60$ (km). Once again the system is strongly C-efficient with respect to a linear cost function. The aggregated trip distribution matrix is

$$\mathbf{T}_{60} = \begin{bmatrix} 7239 & 2761 \\ 2761 & 7239 \end{bmatrix} \tag{4.3}$$

which is consistent with a parameter $\beta = 0.0161$. A prediction based on the parameter $\beta = 0.0137$ calibrated from the original system, is seriously biased. If this parameter is used to predict the system at $d = 60$, the predicted values are

$$\mathbf{T}_{60} = \begin{bmatrix} 6879 & 3121 \\ 3121 & 6879 \end{bmatrix} \tag{4.4}$$

Inspection of (4.1-3) shows that an approach based on a constant β overestimates the change by 200%. If we differentiate the exponent $f[d] = \beta[d] \cdot d$ in the standard gravity model, we get

$$f'[d] = \beta'[d] \cdot d + \beta[d] \tag{4.5}$$

The hypothesis $\beta'[d] = 0$, yields $f'[d] = \beta[d]$. Figure 1 shows the fraction $\beta[d]/f'[d]$ for the above system. Whenever this fraction is close to one, we can expect that a replication based on constancy of the β-parameter will be fairly accurate.

This is OK if $d \approx 25$ (km). If $d > 60$, however, this fraction is above 2, suggesting that a change in the β-parameter is the dominant effect. Thus ignorance of the term $\beta'[d] \cdot d$ leads to a severely biased result. In fact, the original observation T_{80} will be a better prediction of T_{60} in this case!

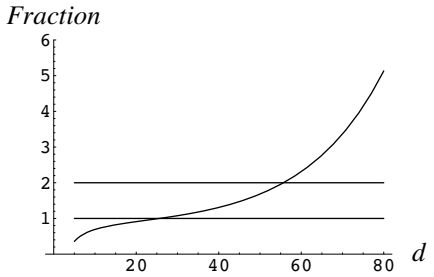

FIGURE 1: The fraction $\beta[d]/f'[d]$ in Example 4.1

Note that the systems at $d = 60$ and at $d = 80$ are both strongly C-efficient and consistent with random utility theory. Hence there is no conflict with any of these theories. The problem lies in the non-constancy of the β-parameter. To examine this further, we show the graph of the mapping $d \mapsto \beta[d]$, see Figure 2.

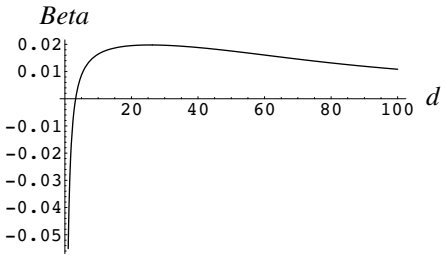

FIGURE 2: Variation of the β-parameter in Example 4.1

It is interesting to notice that β is negative if $d \leq 3$. Hence if $d \leq 3$, the aggregate system is neither C-efficient nor consistent with random utility theory.

A weakness of the above result is that the underlying subcategories are assumed to comply with a constant β-parameter. This in itself may not be a reasonable hypothesis. In the following example, however, no such assumptions are needed.

EXAMPLE 4.2

The next example is similar to the above in that we again consider two towns and non-interacting population groups within these populations. Here, however, commuting is determined from the result of a simplistic behavioral rule: Within each population group the wages in the two towns are different. A worker first applies for the work with the highest wages net of commuting costs. If he don´t succeed, he will be employed in the less favorable position. We assume that qualifications

within each separate category are equally distributed between the two towns, so that the number of successful applicants can be determined from random choice.

As the distance between the towns changes, the preferred alternative will be subject to change. A once favorable position in the opposite town, will be rendered unfavorable when the commuting cost gets too high. Consider a system of N such categories. To carry out a numerical simulation of a system of this sort, we draw wages and sizes of population groups from certain random distributions. A detailed specification of these distributions is unimportant, and is omitted. Once drawn, the quantities are fixed throughout the simulation. For each value of $d = d_{12}$, we can then determine the trip distribution matrix from the behavioral rule. The aggregate system is generally strongly C-efficient, and for each d we can find a parameter $\beta = \beta[d]$ which yields a perfect replication of the system. In Figure 3 we show the result of a numerical simulation with 500 different categories.

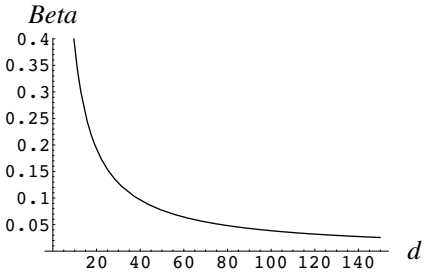

FIGURE 3: Variation of the β-parameter in Example 4.2

As is clearly seen in Figure 3, the β-parameter is subject to a significant change. The change, however, is the result of some specific choice of distributions for the wages and the different population groups. In particular there is no reason to claim that this specific choice is more reasonable than anything else. For the sake of argument let us introduce the additional hypothesis that β is constant, and ask what extra conditions does this impose on the wage distribution? We let $c = c[d]$ denote the cost of commuting if the distance is d. Loosely speaking, the value of β will be reduced at $d = d_0$ if there are too many firms with wage differences in the interval $[c[d_0], c[d_0 + \Delta d_0]]$, it will increase if there are too few. Hence if we assume that the population categories are fixed, constancy of the β-parameter imposes an extremely restrictive condition on the distribution of wage differences.

A closer examination reveals that the intuitive idea above above is a bit too simple. It is true that the result depends on the number of firms, but it also depends on the relative positioning of the population groups. A more rigorous argument can be described as follows: Let E_j denote the total number of working opportunities in category j, and L_j denote the total number of workers. We always assume $L_j = E_j$. Correspondingly, we let E_{ij}, L_{ij} denote the number of working opportunities/workers in town i and category j, respectively. For each category $j = 1, \ldots, N$

we can define an impact factor I_j as follows

$$I_j = L_j \cdot \min\left[4 \cdot \frac{E_{1j}}{E_j} \cdot \frac{L_{2j}}{L_j}, 4 \cdot \frac{E_{2j}}{E_j} \cdot \frac{L_{1j}}{L_j}\right] \tag{4.6}$$

The impact is thus determined by the total number of workers L_j in the category, adjusted by a factor measuring a spread between the two towns. The impact adjustment has a maximal value equal to 1 obtained in the most symmetric state; $E_{1j} = E_{2j} = L_{1j} = L_{2j}$. It is zero if a category has either zero working opportunities or zero workers in one of the towns. Such categories have no impact since the resulting trip distribution is trivial in this case. We define the relative impact R_j by

$$R_j = \frac{I_j}{\sum_{i=1}^{N} I_i} \tag{4.7}$$

The usefulness of the above definitions will be clear from the following theorem:

THEOREM 4.3

Let ΔW_j denote the wage differences in category j, let R_j denote the relative impact of category j as defined in (4.6-7), and let c denote the traveling cost pr. km. If the number of employment opportunities/workers in every town and category is fixed, there exists a smooth function g s.t. the β-parameter for the system in Example 4.2 is approximately constant if and only if

$$\sum_{j: cd < \Delta W_j \le c(d+\Delta d)} R_j \approx g[\beta d] \cdot \Delta d \tag{4.8}$$

Remarks: The system is discrete, and the resulting trip distribution is piecewise constant in d. Hence equality can only be achieved in the limiting case $N \to \infty$. The function g can be written down by an explicit expression, which does not depend on the wage distribution, see Ubøe (2001). This expression is quite complicated, however, so we have chosen to omit the details.

PROOF

This is a straightforward application of Theorem 7.9 and Proposition 7.10 in Ubøe (2001).

□

From Theorem 4.3 we can hence see that a constant β-parameter can only be obtained if the sum of the relative impact factors of all firms with wage-differences

in the interval $[cd, c(d+\Delta d)]$ equals $g[\beta d] \cdot \Delta d$. It is very hard to imagine that firms should comply with a condition of this sort. For a more refined discussion of aggregate systems of the above type, see Glenn et al. (2001).

5. Alternative dynamic restrictions

A system can be equipped with a dynamic hypothesis in many different ways, and we will now discuss some particular features of the examples below:

- β is constant
- The entropy; $H = -\sum_{i,j} T_{ij} \cdot \ln[T_{ij}]$ is constant
- The total cost; $C = \sum_{i,j} c \cdot d_{ij} \cdot T_{ij}$ is constant

These are three different examples of a dynamic hypothesis. There are of course many more. The three quantities are related, and the following expression applies:

$$\frac{\partial H}{\partial \beta} = \beta \cdot \frac{\partial C}{\partial \beta} \tag{5.1}$$

see, e.g., Erlander and Stewart (1978). Let us look at a few simple examples to see what is going on. Consider the 5-node network in Figure 4. In this case we assume that there is only one category of workers, and that the initial system can be described by a standard gravity model with parameter $\beta = 0.03$.

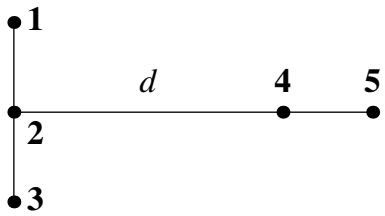

FIGURE 4: A 5-node network

We will assume that $d_{12} = d_{23} = d_{45} = 20$ (km), and that $d_{24} = d$ (km), and that:

$$L_1 = 1000, L_2 = 1000, L_3 = 1000, L_4 = 5000, L_5 = 2000$$

$$E_1 = 1500, E_2 = 2500, E_3 = 1500, E_4 = 3000, E_5 = 1500$$

Initially we assume that $d = 60$. The system is completely determined from the specifications above, and we have

$$\mathbf{T}_{\text{initial}} = \begin{bmatrix} 498 & 326 & 150 & 19 & 6 \\ 230 & 500 & 230 & 30 & 10 \\ 150 & 326 & 498 & 19 & 6 \\ 479 & 1039 & 479 & 2260 & 743 \\ 142 & 309 & 142 & 672 & 734 \end{bmatrix} \tag{5.2}$$

We now introduce a change in the network, reducing the distance d from 60 km to 40 km, and ask what is the effect of the change in the three cases mentioned above.

Case 1: We assume that β is constant. If so, the resulting trip distribution is given by the matrix $\mathbf{T}_\beta$ below.

$$\mathbf{T}_\beta = \begin{bmatrix} 476 & 307 & 143 & 55 & 18 \\ 215 & 460 & 215 & 82 & 27 \\ 143 & 307 & 476 & 55 & 18 \\ 512 & 1096 & 512 & 2160 & 720 \\ 154 & 329 & 154 & 648 & 716 \end{bmatrix} \tag{5.3}$$

Case 2: We assume that the entropy is constant. This determines a unique value of $\beta = 0.0365$ in the final network, and the resulting trip distribution is given by the matrix $\mathbf{T}_{\text{entropy}}$ below.

$$\mathbf{T}_{\text{entropy}} = \begin{bmatrix} 542 & 291 & 126 & 32 & 9 \\ 216 & 498 & 216 & 54 & 16 \\ 126 & 291 & 542 & 32 & 9 \\ 482 & 1112 & 482 & 2256 & 667 \\ 134 & 309 & 134 & 626 & 798 \end{bmatrix} \tag{5.4}$$

Case 3: We assume that the total cost is constant. This determines a unique value of $\beta = 0.0193$ in the final network, and the resulting trip distribution is given by the matrix $\mathbf{T}_{\text{cost}}$ below.

$$\mathbf{T}_{\text{cost}} = \begin{bmatrix} 356 & 314 & 164 & 119 & 47 \\ 203 & 388 & 203 & 147 & 58 \\ 164 & 314 & 356 & 119 & 47 \\ 580 & 1108 & 580 & 1954 & 778 \\ 197 & 375 & 197 & 662 & 570 \end{bmatrix} \tag{5.5}$$

From the results in (5.3)-(5.5) we see that if the entropy is constant, the parameter β increases, and we get less commuting than if β is constant. If on the other hand the average traveling cost is constant, the parameter β decreases, and we get more commuting.

When we fix the distances in the system, each level of the cost defines a unique β value which in turn defines a unique entropy. Hence we can define a function $H = H(C)$. In Figure 5 we show numerical plots of this function for the initial

network, labeled $d = 60$, and for the final network, labeled $d = 40$. From (5.1) we see that

$$\beta = \frac{\partial H}{\partial C} \tag{5.6}$$

Hence the β parameter can be interpreted as the slope of the graphs in Figure 5. In Figure 5 we see that a constant entropy corresponds to point I, and that a constant cost corresponds to point III. If β is constant, the tangents must have the same slope, corresponding to point II in the figure. Clearly this defines three different states of the system.

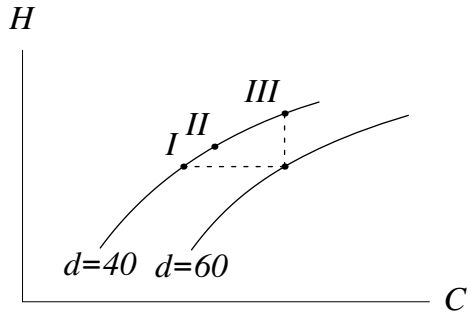

FIGURE 5: Entropy H versus cost C

The important question is now the following: Which of these cases is the "right" solution to the problem? The answer depends crucially on the context, and as we already have seen in Section 4, the answer may well be: None of the above.

Constant cost of commuting is a natural budget constraint, and it is easy to imagine a scenario where this is indeed the governing force in the system. If the entropy is constant, however, certain forces must be present which slow down the response in the system.

Consider the following extreme: All the workers within a job category reside in the same zone, while all the employment opportunities are in a different zone. This puts a forcing restriction on the commuting flows; no matter how the distances change, this particular flow will remain constant. If all the links are reduced by the same percentage, constant entropy is equivalent to constant trip distributions. Hence if the system is dominated by binding flows, constant entropy may in fact be a favorable assumption, at least in some special cases.

What about a constant β parameter? From a random utility approach the parameter β can be interpreted as the marginal utility related to variations in commuting distance. This interpretation is not valid, however, if commuting behaviour is influenced by other distance-dependent attributes related to the spatial configuration

of central places. To be more precise commuting behaviour can be influenced for example by competition or agglomeration effects related to the spatial accessibility of alternative destinations. In addition, however, the observed commuting flow pattern also depends on how separate categories of jobs and workers are distributed between the central places in the geography. Such kind of aggregation problems contribute to make the assumption of a constant distance deterrence parameter inappropriate for making predictions in time and space.

6. Concluding remarks

The problems discussed in this paper have a common core. *C*-efficiency and random utility theory are excellent tools for the discussion of systems that are homogeneous and where the distances are fixed, whereas the replication/prediction problem deals with changes in the underlying structure. The important question in the replication/prediction problem is not how preferences are distributed but rather how preferences change when the system changes. To study this kind of problem, it is necessary to equip the model with an extra behavioral hypothesis. Assuming that the parameters are fixed, is one such extra hypothesis, but as we have just seen, this need not always be a satisfactory solution. It would be much more satisfactory if one could postulate the distance dynamics from some sort of differential principle. The problem with the gravity approach is that this setup is not all that suitable for extensions of this kind. This motivates the search for alternative approaches. Such approaches migth lead to models that are inferior with respect to explanatary power, but for many purposes is is more important to consider the predictability of the model.

REFERENCES

Baxter, M. 1983. "Model specification and spatial structure in spatial-interaction models", *Environment and Planning A*, vol. 15.

Desta, E: and B. WM. Pigozzi. 1991. "Further experiments with spatial structure measures in gravity models". *Tijdschrift voor Econ. En Soc. Geografie* 82, No. 3.

Erlander, S., and T. Smith. 1990. "General representation theorems for efficient population behavior", *Applied Mathematics and Computation*, 36, 173-217.

Erlander, S., and N. F. Stewart. 1978. "Interactivity, accessibility and cost in trip distribution", *Transportation Research*, vol. 12, 291-293.

Erlander, S., and N. F. Stewart. 1990. *The gravity model in transportation analysis - theory and extensions*, VSP, Utrecht.

Ewing, G. 1986. "Spatial pattern in distance-deterrence parameters and Fotheringham's theory of competing destinations", *Environment and Planning A*, vol. 18, 547-551.

Fik, T.J. and G.F. Mulligan, 1990. "Spatial flows and competing central places: towards a general theory of hierarchical interaction", *Environment and Planning A*, vol. 22, 527-549.

Fik, T.J., R.G. Amey and G.F. Mulligan, 1992. "Labor migration amongst hierarchically competing and intervening origins and destinations", *Environment and Planning A*, vol. 22, 527-549.

Fotheringham, A. S. 1981. "Spatial structure and distance-decay Parameters", *Annals of the Association of American Geographers*, vol. 71, no. 3, 425–436.

Fotheringham, A.S. 1983a. "Some theoretical aspects of destination choice and their relevance to production-constrained gravity models". *Environment and Planning A*, vol. 15, 1121-1132.

Fotheringham, A.S. 1983b. "A new set of spatial interaction models: the theory of competing destinations". *Environment and Planning A*, vol. 15, 15-36.

Fotheringham, A.S. 1984. "Spatial flows and spatial patterns". *Environment and Planning A*, vol. 16, 529-543.

Fotheringham, A.S. 1986. "Modeling hierarchical destination choice" . *Environment and Planning A*, vol. 18, 401-418.

Fotheringham, A.S. 1988. "Consumer store choice and choice-set definition" . *Marketing Science* 7, 299-310.

Glenn, P., I. Thorsen, and J. Ubøe. 2001. "A Microeconomic Approach to Distance Deterrence Functions in Modeling Journeys to Work", NHH working paper 12/2001 ISSN: 1500-4066.

Gordon, I.R. 1985. "Economic explanations of spatial variation in distance deterrence", Environment and Planning A, vol. 17, 59-72.

Ishikawa, Y. 1987. "An empirical study of the competing destinations model using Japanese interaction data". Environment and Planning A, vol. 19, 1359-1373.

Lo, L. 1991. "Substitutability, spatial structure, and spatial interaction", Geographical Analysis 23, 133-146.

Manski, C.F. 1977. "The structure of random utility models", Theory and Decision 8, 229-254.

Pellegrini, P.A., A.S. Fotheringham and G. Lin. 1997. "An empirical evaluation of parameter sensitivity to choice set definition in shopping destination choice models", Papers in Regional Science 76, 257-284.

Sen, A. and T. Smith. 1995. "Gravity models of spatial interaction behavior", Springer-Verlag Berlin Heidelberg.

Thill, J-C. 1992. "Choice set formation for destination choice modeling", *Progress in Human Geography* 16, 361-382.

Thill, J-C and J. L. Horowitz. 1997. "Travel-time constraints on destination-choice sets", *Geographical Analysis* 29, 108-123.

Thorsen, I. and J. P. Gitlesen. 1998. "Empirical evaluation of alternative model specifications to predict commuting flows", *Journal of Regional Science*, 38, 273-292.

Ubøe, J. 2001. "Aggregation of Gravity Models for Journeys-to-work", NHH working paper 4/2001 ISSN: 1500-4066.