URBAN SYSTEMS MODELING

Piotr Korcelli, Editor
International Institute for Applied Systems Analysis, Laxenburg, Austria

RR-83-21
August 1983

Reprinted from *Sistemi Urbani*, volume 4 (1982)

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
Laxenburg, Austria

Reprinted with permission from *Sistemi Urbani* 4(3):251–483, 1982.

FOREWORD

Many large urban centers in developed nations have experienced population decline and most have grown at rates lower than those of middle-sized and small settlements. Current patterns of urban contraction are generating fiscal pressures at the local level and may also have important consequences for economic development at the national level. Such trends constitute a serious challenge to spatial policy analysts, whose traditional settlement policies are no longer considered adequate.

The objective of the Urban Change Task in IIASA's former Human Settlements and Services Area was to synthesize available empirical and theoretical information on the principal determinants and consequences of recent patterns of growth and decline in developed countries. Investigations were made of (a) the role played by declining rates of national population growth and changing migration propensities in urban change processes, (b) interdependences between industrial restructuring at a national level and adjustments observed in local urban economies, and (c) relations between aggregate patterns of urban development and their impact on the evolution of the internal spatial structure of urban areas.

The papers in this special issue focus primarily on the third aspect of urban development, one that provided linkages between the Urban Change and the Public Facility Location Tasks within the Human Settlements and Services Area. The papers are selected, revised contributions to the Workshop on Urban Systems Modeling held in Moscow in September–October 1980. The articles consider a number of issues, specific variables, and general techniques that are currently used by urban modelers. The more substantive aspects include the role of alternative housing allocation policies, as well as environmental factors in the process of urban change.

ANDREI ROGERS
Chairman
of the former
Human Settlements and Services Area

CONTENTS

Indice

Sistemi Urbani, 3, 1982, 251-252 Guida Editori, Napoli, Italia

Introduction

Disputes concerning conceptual virtues and limitations of urban models have acquired a decade-long tradition (Goldner, 1971; Lee, 1973; Sayer, 1976; Batty, 1979a, 1979b). At the same time, the need for further development of such models becomes ever greater. This demand, generated in the past by problems of urban management at a local level, is now strengthened in view of recent urban trends which involve a broad spectrum of patterns, from absolute population decline in some of the large urban agglomerations in highly urbanized countries to the accelerating concentration of population and economic activity in the primate cities of many developing countries. As a consequence of these divergent trends, planners are now facing necessity of reformulating settlement policies carried out at the national and urban levels, since many of the existing policies have either supported trends no longer deemed favorable or have failed to produce more balanced settlement patterns.

Internal functioning of urban areas represents one of the substantial facets as well as determinants of settlement change at a national scale. Inefficient spatial organization, as expressed for example in high friction to travel and relocation within urban space, bears heavily on overall quality of life in urban agglomerations. This may in turn result in a long-term relative decline of the position of large cities vis-a-vis other components of settlement systems. Owing to the role of large cities as seedbeds for technical and organizational innovations, such trends may be negatively evaluted from the point of view of economic development at a national level (Richardson, 1978).

Admittedly, existing urban policy-oriented models say little about how and why economic activities originate within, or migrate to and out of an urban area, and what the likely consequences such moves may have on other urban areas. The list of variables and interdependencies which are crucial for the understanding or urban functioning and change, but largely neglected in the available modeling frameworks, is a rather long one. It includes, among others, various aspects of population dynamics (such as changing household composition and functions), land-use conversion patterns, substitution among alternative types of movement (such as migration and daily travel), evolution of the functions of distance. Also, methods for the evaluation of urban performance, and of spatial organization of urban areas, are not well developed.

The papers included in this volume attempt to address a number of current urban modeling issues. Some contributions focus on problems found on the interface between urban planning and urban modeling, while others propose more efficient computational algorithms. Still other papers take a critical view on methods and approaches which have so far been dominant in the field. On the substantive side, the topics vary from urban transportation systems to housing allocation, the urban environment and general planning theory. Many readers may find the list of approaches and themes followed by different authors to be excessively large. It seems, however, that such a diversity represents an essential condition for further progress to be achieved in urban modeling.

This issue of the journal is composed of selected contributions to the Workshop on Urban Systems Modeling, held in Moscow during September 30 - October 4, 1980. The Workshop was organized jointly by the Committee for Systems Analysis of the Presidium of the USSR Academy of Sciences (through its Scientific Council of Management of the Development and Functioning of Towns and Settlement Systems), the Human Settlements and Services Area of the International Institute for Applied Systems Analysis (Laxenburg, Austria), and the International Scientific Research Institute of Management Problems (Moscow).

P. Korcelli
Human Settlements and Services Area
IIASA

References

Batty M. (1979a) Paradoxes of Science in Public Policy: the Baffling Case of Land Use Models, *Sistemi Urbani, 1*, 1, 89-122.

Batty M. (1979b) Progress, Success and Failure in Urban Modeling, *Environment and Planning A, 11*, 863-878.

Goldner W. (1971) The Lowry Model Heritage, *Journal of the American Institute of Planners, 37*, 100-110.

Lee D. Jr. (1973) Requiem for Large-Scale Models, *Journal of the American Institute of Planners, 39*, 163-178.

Richardson W. (1978) Basic Economic Activities in Metropolis, in Leven C.L. (ed.) *The Mature Metropolis*, Heath and Co., Lexington, Mass., 251-280.

Sayer A. (1976) A Critique of Urban Modelling: from Regional Science to Urban and Regional Political Economy, *Progress in Planning, 6*, 187-254.

Sistemi Urbani, 3, 1982, 253-266 Guida Editori, Napoli, Italia

Principles of formulation and analysis of urban models

Yu. S. Popkov, B.L. Shmulyian
Institute for Systems Studies, 29 Ryleyev Street, 119034 Moscow, U.S.S.R.

Received 30 September 1980, in revised form 12 March 1981

Abstract. The methodology discussed in the paper is based on the assumption that an urban area can be represented as a two-level system. Within such a system the links between micro-level elements are random in character while links between states at the macro-level are of deterministic nature. Interaction of the two levels results in the transformation of the random moves into a regular process. The micro-level is formed by an ensemble of choice of facilities by individual residents described by probability distribution functions. The macro-level is composed of functionally homogenous subsystems such as workplaces, housing stock, population characteristics, etc. The integration of different subsystems is achieved using entropy-maximization functions. Methods of studying the models' stability are proposed.

Key words: urban system, behavioral models, entropy maximization, equilibrium-stable state.

1. Introduction

The process of intensive urbanization taking place in the world of today is accompanied by the growth of existing cities, and the formation of new cities and urban agglomerations. This process is characterized by a number of problems and negative externalities related to environmental degradation, deterioration of accessibility within large urban systems, difficulties of controlling the spread of urban areas and so forth.

Therefore the problem of an efficient organization of urban areas appears to be very important and essential. At least two approaches to solving this problem may be noted. The first comprises construction of principles for optimal planning of an urban area. Here the definition of precise quantitative criteria to evaluate the quality of allocation alternatives is implied. A second approach (let us call it behavioral) attempts to develop an areal plan that takes into account the behavior of residents using facilities allocated in this area.

Practical realization of this latter approach comprises construction of model simulating residents' behavior in using these various facilities. The Lowry model (1964), based on a gravity-based description of residents' behavior, was the first major model of this class. This approach was further developed in the works of Wilson (1967, 1974) and other authors. In his models Wilson assumes that a resident chooses a facility in random way and independently of other residents' behavior. This assumption leads to schemes analogous to the distribution of particles over energy levels whose properties are well known from statistical physics. The stable state of such systems is reached when their entropy is maximal. We note that as a rule Wilson's models describe a procedure of choosing between functionally homogeneous objects while in the Lowry model much attention is paid to description of interactions between functionally heterogeneous facilities such as the basic sector, and service and residential sector facilities.

In this paper a methodology for constructing models of human settlement is laid down. These models make it possible to determine the functional-spatial structure of urban areas by simulating human behavior when choosing facilities to interact with. We shall refer to these as functional-spatial models (FSM). In these models main attention is paid to describing interactions between urban subsystems.

2. General structure of a functional-spatial model of human settlement systems

Let us consider a closed (autonomous) settlement system at fixed moment of time, or during a time interval when its state does not vary. A city, or a number of cities sufficiently closely related to each other, may serve as an example of such a system. We consider the area occupied by this system and divide it into M regions. Facilities of different functional types are allocated in each region.

We denote: d^{μ}_{m} - a functional facility of type μ located in district $m \in \overline{1, M}$; q - the number of functional types, i.e. $\mu \in \overline{1, q}$; and $N = \{N_1, \dots, N_M\}$ - the allocation of population over the urban area.

Residents patronise these facilities d^{μ} for consumption purposes. It appears natural to assume that the procedure of individual choice of some or another functional facility by an individual for consumption purposes is stochastic.

A quantitative model of this procedure is constructed from characteristics of some individual n_l, the facilities d^{μ}_{m} and the transportation media facilitating the interaction.

Each object d^{μ}_{m} is characterized by a set of attributes $\{\sigma^{\mu}_{1}, \dots, \sigma^{\mu}_{r_\mu}\} = \sigma^{\mu}$, that are independent of the location of the facility. The distribution function F^{ν}, determining the probability of choosing facility d^{μ}_{m} by resident n_l located in region l is our quantitative description of behavior. Since F^{ν} characterizes a pairwise interaction we obtain:

$$F^{\nu} = F^{\nu}(n^{\nu}_{l}, d^{\mu}_{m}) .$$

It is necessary to note that the procedure of choice is very complex, and not investigated in full here. Many assumptions are made pertaining to the mechanism of choice. In particular, one of them (perhaps the most widespread) states that probability F is a function of distance ρ between d^{μ}_{m} and n_{l}.

Because of their multitude, individual stochastic interactions between n_l and d^{μ}_{m} generate in the aggregate completely deterministic characteristics of the system's behavior. These are a population allocation N; attributes $\{e^{\mu}_{1},\ldots,e^{\mu}_{r_{\mu}}\} = \tilde{E}^{\mu}$, characterizing groups of functionally homogeneous facilities (subsystems D^{μ}), or the distribution of these attributes $E^{\mu} = \{E^{\mu}_{1},\ldots,E^{\mu}_{m}\}$ over the set of locations.

Two interconnected levels are distinguished in the functional-spatial model (FSM) of a human settlement system. At the first level, the *microlevel*, stochastic individual interactions between population elements n_l and facilities d^{μ}_{m} from different functional subsystems are taking place. The second, the *macrolevel*, gives a transformation of stochastic interactions into a deterministic state, characterized by subsystems vectors $E^{1},\ldots,E^{q}$ and a vector of population allocation N.

3. FSM macrolevel

The macrolevel is constructed from subsystems D^{μ} ($\mu \in \overline{1,q}$), combining functionally homogeneous facilities d^{μ}. Each subsystem is characterized by the vector index E^{μ} ($\mu \in \overline{1,q}$). In addition to these vectors, the macrolevel state (hence the model's state) is completed by a characterization of the vector of population allocation N. The main mechanism of the system, namely residential choices of facilities from subsystems D^{μ}, is defined as a correspondence between E^{μ} and N: $E^{\mu} = E^{\mu}(N)$ for all $\mu \in \overline{1,q}$.

To form the model it is necessary to link subsystems D^{μ} ($\mu \in \overline{1,q}$) in some way. The simplest linkage is a hierarchical chain whereby different types of facilities are located sequentially. But it is impossible to establish a «rigid» hierarchy of subsystems in a human settlement system. This is explained by its essential heterogeneity, because ecological, socialdemographic, technical, economic and other factors influence the system within essentially a bounded area. Therefore any traditional orderings of the subsystem: for instance basic sector, housing, service, population, transport - may be preferred to others at different stages in the system's development.

As a result any subsystem may occupy any level of the hierarchical chain in a random way. In this case formulation of an urban system model may be realized in Popkov (1976). The idea is to define a hierarchical chain which is the most informative representative of some stochastic hierarchical chains ensemble. This chain then describes a sequence of subsystems which may be represented using subsystems states indexed as:

$$E^{\mu_1^*} > W^{\mu_2^*} > \ldots > E^{\mu_q^*} \;. \tag{1}$$

Below we shall renumber the subsystems such that $\mu_j^* \equiv j$.

We recall that each index E^μ is a function of vector N of population distribution. Then, under a hierarchical sequencing of location problems, we obtain:

$$\begin{aligned} E^1 &= T_1(N) \\ E^2 &= T_2(E^1,N) \\ &\cdots\cdots \\ E^q &= T_q(E^{q-1},N) \end{aligned} \tag{2}$$

where T_μ are operators relating index $E^{\mu-1}$ to distribution N and index E^μ of the next hierarchical level.

As noted earlier, the individual choice of a facility d_m^μ by an individual n_l depends on characteristics of n_l and on d_m^μ. Therefore the distribution N depends on all subsystems E^μ, i.e.:

$$N = B(E^1,\dots, E^q) \tag{3}$$

where B is an operator relating N to $E^1,\dots,E^q$. Equations (2), (3) describe a structural formation at a macrolevel of this functional-spatial model of a human settlement system (fig. 1). From equations (2), (3) and the structural scheme (fig. 1) it follows that FSM is a closed system containing the hierarchical chain $T_1,\dots,T_q$ and a feedback B.

This allows us to realize the model on a computer as a closed iterative procedure consisting of the sequential application of operators:

$$\begin{aligned} E^{1(\tau+1)} &= T_1(N^{(\tau)}) \\ E^{2(\tau+1)} &= T_2(E^{1(\tau+1)}, N^2) \\ &\cdots\cdots \\ E^{q(\tau+1)} &= T_q(E^{q-1(\tau+1)}, N^\tau) \\ N^{(\tau+1)} &= B(E^{1(\tau+1)},\dots, E^{q(\tau+1)} , \end{aligned} \tag{4}$$

where $\tau = 0,1,2,\dots$ is the step of the iterative procedure. To determine operators $T_1,\dots,T_q$ and B it is necessary to consider the microlevel of this model.

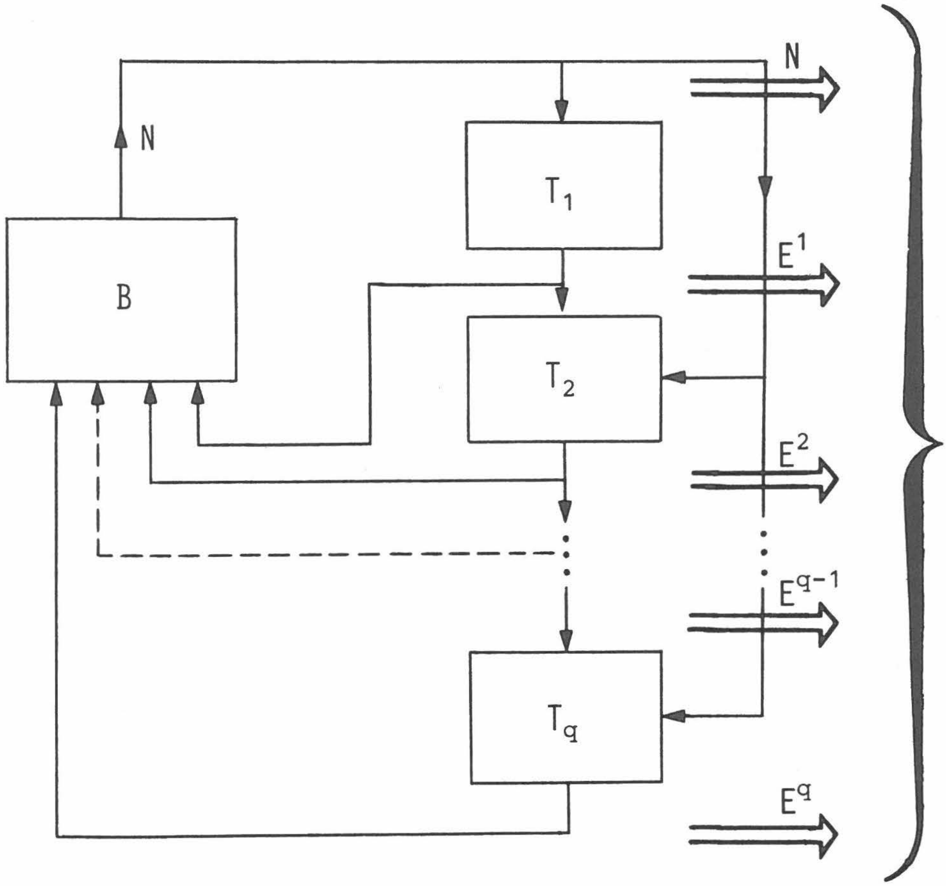


Figure 1 Macro-structure of an FSM model

4. FSM microlevel

The FSM microlevel represents a formal description of stochastic individual pairwise choice of facilities d^{μ}_{m} from functional subsystems D^{μ} ($\mu \in \overline{1,q}$) by individuals n_l.

It is clear that individual choice $\{n_l \rightarrow d^{\mu}_{m}\}$ depends on a number of social, economic, demographic, psychological, status, and other factors. Nevertheless, assuming a stochastic perspective on individual choice, this may be generally characterized by some probability distribution function F depending on n_l and d^{μ}_{m}. Decomposition of this relation allows us to distinguish three groups of factors completely determining the distribution $F(n_l, d^{\mu}_{m})$:

- *behavioral factors*, characterizing an individual n_l (social and demographic attributes, economic status, individual priorities and perceptions of the quality of consumption facilities and means of transportation);
- *generalized quality factors* of chosen facilities d^{μ}_{m} and their localization regions m (places of work: professional composition, wage level, prestige, general production level; and service facilities: service quality, baskets of goods offered, time spent for service, etc.);
- *communication factors*, characterizing the environment where individual choice is made (distance, generalized travel cost, etc.).

According to these groups of factors we define three probability distribution functions:

$f(l \mid d^{\mu}_{m}, h_{ml})$ a conditional probability distribution of choice of a fixed facility d^{μ}_{m} by an individual n_l when the characteristic h_{ml} of transportation networks is fixed;

$\varphi^{\mu}(m)$ a probability distribution of generalized quality of facilities d^{μ}_{m} over the urban area;

$w^{\mu}(l,m)$ a probability distribution of characteristics of transportation and distance between l and m.

We suppose that the two last distributions are mutually independent.

Then the probability distribution of individual choice is

$$F^{\mu}(l,m) = f(l \mid d^{\mu}_{m}, h_{ml})\, \varphi^{\mu}(m)\, w^{\mu}(l,m) \ . \tag{5}$$

When realizing individual choices among any pair (l,m) a stochastic flow x^{μ}_{lm} of individual choices is constructed. This results in a flows matrix X^{μ}, which is stochastic in the sense that there is an ensemble of such matrices defined for some probability distribution function.

Assuming that individual choices are independent of one another and using traditional assumptions (Landau, Lifshetz, 1964; Wilson, 1974;

Imelbayev, Shmulyian, 1977), it may be shown that

$$\mathring{X}^{\mu} = \arg\max H^{\mu} \tag{6}$$

where $H^{\mu} = \sum_{l,m=1}^{M} x^{\mu}_{lm} \ln \frac{F^{\mu}(l,m)}{x^{\mu}_{lm}}$ is the generalized entropy of the distribution $\{x^{\mu}_{lm}\}$.

If we also assume that the human settlement system under consideration is closed it follows that the stochastic flows x^{μ}_{lm} must satisfy the following constraints for population:

$$\sum_{m=1}^{M} x^{\mu}_{lm} = C^{\mu} N_l , \qquad l \in \overline{1,M} \tag{7}$$

and for limiting values, G^{μ}_m, of subsystems attributes, E^{μ}_m:

$$\sum_{l=1}^{M} x^{\mu}_{lm} \leq G^{\mu}_m \tag{8}$$

where C^{μ} is the coefficient characterizing the share of residents choosing facilities from subsystem μ. Here it is assumed that E^{μ}_m is a scalar index; if it is vector, then analogous constraints are formulated component-wise.

Therefore the problem of generalized entropy maximization (6) under constraints (7, 8) describes the process of forming a FSM microlevel state from stochastic flows x^{μ}_{lm}, which are in turn, the results of individual choice.

Therefore operators $T_1, \ldots, T_q$ of the hierarchical chain in the FSM structural scheme are

$$T_1 : \left[E^1_m = \sum_l \mathring{x}^1_{lm} , \; m \in \overline{1,M} \; ; \right.$$

$$\left. \mathring{x}^1_{lm} = \arg\max \left\{ \sum_{lm} x^1_{lm} \ln \frac{F^1(l,m)}{x^1_{lm}} \;\middle|\; \sum_m x^1_{lm} = C^1 N_l, \; l \in \overline{1,M} \right\} \right] \tag{9}$$

$$T_\mu : \Big[E^\mu = \sum_l \mathring{x}^\mu_{lm}, \quad m \in \overline{1,M} \; ;$$

$$\mathring{x}^\mu_{lm} = \arg\max \left\{ \sum_{lm} x^\mu_{lm} \ln \frac{F^\mu(l,m)}{x^\mu_{lm}} \;\middle|\; \sum_m x^\mu_{lm} = C^\mu N_l, \; l \in \overline{1,M} \; ; \right. \tag{10}$$

$$\left. \sum_l x^\mu_{lm} \le \alpha^\mu E^{\mu-1}_m + \beta^\mu; \; m \in \overline{1,m} \right\} \Big]$$

where G^μ_m is a linear function of the preceding subsystem indices $E^{\mu-1}_m$.

In this FSM structural scheme, operators B characterize the process of population allocation under the assumption that subsystem attributes $E^1, \ldots, E^q$ are known. According to this scheme (fig. 1) allocation N is formed from the stochastic flows y_{lm}. Subsystems D^μ ($\mu \in \overline{1,q}$) are sources of these flows. When the flows are mutually independent the block B's state will be described by the three-dimensional matrix

$$\mathring{Y} = \arg\max H \tag{11}$$

where

$$H = \sum_{\mu=1}^{q} \sum_{l,m=1}^{M} y^\mu_{lm} \ln \frac{\tilde{F}^\mu(l,m)}{y^\mu_{lm}} \; . \tag{12}$$

Elements of $\mathring{Y}$ must satisfy constraints analogous to (8) for all $\mu \in \overline{1,q}$.

Operator B is then of the form:

$$B_\mu : \Big[N_l = \sum_\mu \sum_m \mathring{y}_{lm}, \quad l \in \overline{1,M} \; ;$$

$$\mathring{y}^\mu_{lm} = \arg\max \left\{ \sum_\mu \sum_{lm} y^\mu_{lm} \ln \frac{\tilde{F}^\mu(l,m)}{y^\mu_{lm}} \;\middle|\; \sum_l y^\mu_{lm} = \gamma^\mu E^\mu_m \; , \right. \tag{13}$$

$$\left. m \in \overline{1,M}; \; \mu \in \overline{1,q} \right\} \Big] \; .$$

It can thus be seen that the FSM constitutes a sequence of generalized conditional entropy problems relates to each other.

5. Equilibrium - stable state

Consider relations within the problem introduced here, i.e. procedure (4), in more detail. Each operator (9), (10), (13) represents the constrained entropy maximization problem. Therefore we may consider the stable states $\{\mathring{y}^{\mu}_{lm}\}$ or $\{\mathring{x}^{\mu}_{lm}\}$ realized in each subsystem. These states depend on subsystem parameters, including the right-hand sides of constraints. Since the right sides represent other subsystems indices E^{μ}_{m}, N_l, stable states of each subsystem depend on other subsystems' stable states. Therefore the concept of a subsystem's stable state does not suffice for determination of a «natural» state of the system.

So we have to use the concept of «equilibrium-stable» state of the system $U^* = \{x^{\mu*}_{lm}, y^{\mu*}_{lm}\}$, which has the following properties:

- it is a stable state for each subsystem;
- it is an equilibrium for all subsystem, i.e. U^* transforms all constraints in the operators T_i, B into identities.

As an example of a physical system with such properties consider several cylinders filled with gas, with the cylinder pistons linked together by a leverage. A stable state for each cylinder corresponds to the entropy maximizing state of gas filling this cylinder under the constraint determined by the piston position (accounting for other factors, such as temperature and so on). Changes in constraints on feasible states in any subsystem (change in piston's position in a cylinder) lead to changes in constraints on feasible states in other subsystems (cylinders).

An equilibrium-stable state of this cylinder system is characterized by the stable state of each piston and mechanical equilibrium of the piston system. If a constraint is changed instantaneously (for example, by changing a piston-rod length), the whole system will begin to change its state and after some transient process a new equilibrium-stable state will be realized.

A similar transient process will take place if the initial state is not equilibrium stable. Notice that these transient processes include one more process, namely each subsystem reaching its stable state under fixed external condition. This process is characterized by a relaxation time which is negligible when compared to the times of mechanical moves, as can be shown by physical experiments. So we may assume that these transient processes describe an instantaneously changing stable state of each subsystem.

If we were to construct a mechanical linkage of cylinders in such a way that piston-rods would move sequentially (returning to the first rod after the move of the last), the analogy between the physical system and the process (4) becomes complete. Notice that parameters (such as temperature, pressure, etc.) for the states of each subsystem (cylinder) differ even at the equilibrium-stable state of the system.

Therefore the process (4) describes the urban system reaching its equilibrium-stable state. But convergence to such a state still must be proven. A general approach to investigating such convergence processes is given in Shmulyian (1980a) here we consider only some stages of this study. First of all we notice that stable states of the subsystems, $\mathring{x}^{\mu}{}_{lm}$, $\mathring{y}^{\mu}{}_{lm}$ depend (for given F, $\tilde{F}$) only on subsystem attributes E^{μ}_{m}, N_l. On the other hand, these states $\mathring{x}^{\mu}{}_{lm}$, $\mathring{y}^{\mu}{}_{lm}$ provide intermediary data from which the indices E^{μ}_{m}, N_l are derived. Therefore it is expedient to determine and investigate relations between subsystem μ and other subsystems indices.

For operators T_i with one-sided equality constraints this relation is linear. For example, for T_1 (see Shmulyian, 1980b)

$$E^{1}_{m} = c^{1} \sum_{l} \hat{f}^{1}_{lm} N_l , \qquad m \in \overline{1,M} \tag{14}$$

where $\hat{f}^{1}_{lm}$ are normalized probabilities $F^{1}(l, m)$. For operators with two-sided constraints or with constraints operating as inequalities such relations are not determined explicitly. For example, for operators T_i the solutions are of the form

$$x^{\mu}_{lm} = r^{\mu}_{l} s^{\mu}_{m} F^{\mu}(l, m) \tag{15}$$

where parameters r^{μ}_{l}, s^{μ}_{m} are limiting values of an iteration process which is the «generalized balancing method» (Shmulyian, 1980b):

$$\begin{aligned} r^{w+1}_{l} &= c^{\mu} N_l / \sum_{m} s^{w}_{m} F^{\mu}(l,m) \\ s^{w+1}_{m} &= \min(1, E^{\mu-1}_{m} / \sum_{l} r^{w+1}_{l} F^{\mu}(l,m) \\ s^{0}_{m} &\equiv 1 \; ; \qquad w = 0,\ 1,\ 2, \ldots \end{aligned} \tag{16}$$

(for the variables r_l, s_m index μ is omitted).

Nevertheless, from strict concavity of the objective function (H^{μ}) we conclude that x^{μ}_{lm} (and subsequently E^{μ}_{m}) are continuously determined by the right-hand sides of the constraints. In other words

$$\begin{aligned} E^{\mu}_{m} &= \varphi^{\mu}_{m}(N_l,\ l \in \overline{1,M} \; ; \;\; E^{\mu-1}_{m},\ m \in \overline{1,M}) \\ m &\in \overline{1,M} \; ; \quad \mu \in \overline{2,q} \end{aligned} \tag{17}$$

while for all functions φ bounded variation of the right-hand sides leads to bounded variation of subsystem indices, i.e.

$$|\varphi(Z_1 + \Delta Z_1, \ldots, Z_z + \Delta Z_z) - \varphi(Z_1, \ldots, Z_z)| \leq \sum_t Z_t |\Delta Z_t| . \tag{18}$$

Therefore the general relation between subsystems indices

$$Z = \xi(Z) \tag{19}$$

may be considered as an equation system to determine Z, and the process (4), formulated using indices Z, is then a way to solve this system.

$$Z^{\tau+1} = \xi(Z^{\tau}) . \tag{20}$$

Convergence of the process (4) corresponds to convergence of the iterative process for solving (19). Rewriting the process (20) in variational notation, we obtain:

$$|\Delta Z^{\tau+1}| \leq T |\Delta Z^{\tau}| \tag{21}$$

where T is a matrix with coefficients determined through (17) and linear relations in the form of (14). A sufficient convergence condition is

$$\|T\| < 1 \tag{22}$$

where $\|T\|$ is the determinant of T, from which a set of constraints on parameters c^{μ}, α^{μ}, β^{μ} and so on may be derived.

6. Relation to the entropy maximization problem for the whole system

The concept of the system equilibrium-stable state introduced above allows us to determine [according to the algorithm (4)] a number of «natural» subsystems states $U^* = \{x^{\mu*}_{lm}, y^{\mu*}_{lm}\}$, each of them maximizing its entropy H^{μ}, H. So the question arises - what are the properties of the system's entropy $\overline{H}$? To answer this question we have to define this entropy. It appears natural to define it as the sum of the subsystem entropy functions, i.e.

$$\bar{H} = \sum_{\mu} H^{\mu} + H \tag{23}$$

$$= \sum_{\mu} \sum_{lm} x^{\mu}_{lm} \ln \frac{F^{\mu}(l,m)}{x^{\mu}_{lm}} + \sum_{\mu} \sum_{lm} y^{\mu}_{lm} \ln \frac{\tilde{F}^{\mu}(l,m)}{y^{\mu}_{lm}} \tag{24}$$

for feasible states x^{μ}_{lm}, y^{μ}_{lm} corresponding to all constraints, i.e.

$$\begin{aligned} &\sum_{m} x^{\mu}_{lm} = c^{\mu} \sum_{\mu} \sum_{m} y^{\mu}_{lm}, \quad \mu \in \overline{1,q}\,; \; l \in \overline{1,M} \\ &\sum_{l} x^{\mu}_{lm} \leq \alpha^{\mu} \sum_{l} x^{\mu-1}_{lm} + \beta^{\mu}, \quad m \in \overline{1,M}\,; \; \mu \in \overline{2,q} \\ &\sum_{l} y^{\mu}_{lm} = \gamma^{\mu} \sum_{l} x^{\mu}_{lm}\,; \quad m \in \overline{1,M}\,; \; \mu \in \overline{1,q}\,. \end{aligned} \tag{25}$$

Recall that the entropy of system's state U is the logarithm of the probability of its realization (see Landau, Lifshetz, 1964; Wilson, 1974). From (24) it follows that

$$P(x) = P\{y^{\mu}_{lm}\} \prod_{\mu} P\{x^{\mu}_{lm}\}$$

where $P(x)$ is the probability that system state x will occur. This follows because from (24) the realizations $\{y^{\mu}_{lm}\}$ and $\{x^{\mu}_{lm}\}$, $\mu \in \overline{1,q}$ are independent. In reality, however, subsystem states depend on each other. Thus definition (24) of system entropy is not valid in general.

To illustrate this last point consider once more the example referred to in section 5; i.e. cylinders with mechanical linkages. If we join the cylinders by pipes (allowing the different gases to mix), than after the transient process the stable state corresponding to maximum entropy will be realized. But parameters (temperature, pressure) for each subsystem will be the same and the entropy of the system will be larger than the sum of entropy values in the initial system.

Formally, entropy maximization (24) under constraints (25) is a nonlinear programming problem with a strictly concave objective function and convex feasible set. Its solution $\hat{U} = \{\hat{x}^{\mu}_{lm}, \hat{y}^{\mu}_{lm}\}$ is known to be unique, but to get effective algorithms for solution is rather difficult. On the other hand the equilibrium-stable state $U^{*} = \{x^{\mu *}_{lm}, y^{\mu *}_{lm}\}$, as was already noticed in section 5, corresponds to the (manageable) solution of the system (19) for indices Z. This is a feasible state, but in general it does not coincide with $\hat{U}$. Again notice that the entropy of the system's stable state $\overline{H}(\hat{U})$ is larger than the sum of subsystems entropy values in the equilibrium-stable state, i.e. $\overline{H}(U^*)$.

Finally, from the behavioural point of view the concept of the equilibrium-stable state corresponds to a sequence of independent residential choices of pairs «origins-destinations» (operators T_i, B). Maximization of system's entropy corresponds to the simultaneous independent residential choice of housing and a set of facilities $\mu \in \overline{1,q}$.

7. Conclusion

In this paper main attention is paid to the interaction of urban subsystems. An essential factor for constructing models of interaction is a set of hypotheses on residents' behavior. The assumptions made in the paper appear rather general; variations of quantitative dependencies used here allow us to obtain different models specifying the spatial distribution of the various subsystems in an urban area.

References

Imelbayev S.S., Shmulyian B.L. (1977) Analysis of Stochastic Communication Systems Using Thermodynamical Approach, *Automatica i Telemekhanika, 5*, 77-87 (in Russian).

Landau L.D., Lifshetz E.N. (1964) *Statistical Physics*, Nauka, Moscow (in Russian).

Lowry I.S. (1964) A Model of Metropolis, RM-4035-RC, The Rand Corporation, Santa Monica, California.

Popkov Yu.S. (1976) Information Analysis of Stochastic Hierarchy Systems, Proceedings of VII All-Union Conference of Control Problems, Minsk, 85-92 (in Russian).

Shmulyian B.L. (1980a) Entropy Systems Interaction Models and Application in Spatial Allocation of Urban Subsystems Problems, Paper presented to the Task-Force Meeting «Normative Model Location», IIASA, Laxenburg, Austria.

Shmulyian B.L. (1980b) Spatial Modelling of Urban Systems: an Entropy Approach, Collaborative Paper 80-13, IIASA, Laxenburg, Austria.

Wilson A.G. (1967) A Statistical Theory of Spatial Distribution Models, *Transportation Research, 1*, 253-269.

Wilson A.G. (1974) *Urban and Regional Models in Geography and Planning*, Wiley, New York.

Résumé. La méthodologie mise en cause dans cet essai est basée sur la supposition qu'une zone urbaine peut être représentée par un système à deux niveaux. Dans un tel système les liens entre les éléments du micro-niveau ont un caractère casuel, alors que les liens entre les états au macro-niveau ont une nature déterministe.
L'intéraction des deux niveaux se manifeste par la transformation des mouvements casuels dans un processus régulier. Le micro-niveau est formé d'un ensemble des choix des opportunités des résidents individuels, et est décrit par des fonctions de probabilité de distribution. Le macro-niveau est composé par des sous-systèmes omogènes, du point de vue fonctionnel, tel que les lieux de travail, l'ensemble résidentiel, les caracteristiques de la population. L'intégration des différents sous-systèmes est réalisée en utilisant des fonctions de maximisation de l'entropie. On propose en suite quelques méthodes d'analyse de la stabilité des modèles.

Riassunto. La metodologia discussa in questo articolo è basata sull'assunzione che un'area urbana può essere rappresentata come un sistema a due livelli. Entro un tale sistema, i legami tra gli elementi del microlivello sono di carattere casuale mentre i legami tra gli stati al macrolivello sono di natura deterministica. L'interazione dei due livelli si manifesta nella trasformazione dei movimenti casuali in un processo regolare. Il microlivello è formato da un insieme delle scelte delle opportunità effettuate dai residenti individuali ed è descritto da funzioni di probabilità di distribuzione. Il macrolivello è composto da sottosistemi funzionalmente omogenei, quali i posti di lavoro, lo stock residenziale, le caratteristiche della popolazione ecc. L'integrazione dei diversi sottosistemi viene realizzata utilizzando funzioni di massimizzazione dell'entropia. Si propongono, infine, alcuni metodi di analisi della stabilità dei modelli.

Sistemi Urbani, 3, 1982, 267-310 Guida Editori, Napoli, Italia

A multiactivity location model with accessibility - and congestion - sensitive demand

G. Leonardi

IIASA, International Institute for Applied Systems Analysis, Schlossplatz 1,
A-2361 Laxenburg, Austria.

Received 30 September 1980, in revised form 12 March 1981

Abstract. This paper has two aims. The first one is to build a generalization of the doubly-constrained spatial interaction model, in order to account for sensitiveness of demand to accessibility and congestion and for possible multiple interacting activities. In Section 2 it is shown how this can be done by keeping an extremal representation for the model, which is closely related to the Neuburger's consumer surplus maximizing principle. The second aim is to embed the model developed in Section 2 in a multiactivity optimal location problem, and to develop operational tools to solve the resulting mathematical programming problems. This subject is treated in Section 3. Section 4 is devoted to the discussion of three possible applications: the urban system, the health care system, and the retail system.

Key words: location model, elastic demand, accessibility, congestion.

1. General introduction

Although most location-allocation models deal with a single category of facilities at a time, in real urban areas different facilities, activities, and settlements are present at the same time. Usually interactions take place among them in terms of customers' journeys, exchange of goods, money flows, and information. These interactions tie all activities together, and have to be taken into account in model building of both a descriptive and a normative character.

Many descriptive models of multiactivity systems have been built since the well-known *Model of Metropolis* (Lowry, 1964), and their mathematical formulation and economic interpretation have recently been greatly improved (see MacGill, Wilson, 1979, for a review). The introduction of normative features (i.e., the optimal size and location of physical stocks) also appears in some works (such as Coelho, Williams, 1978; Boyce, LeBlanc, 1979; Leonardi, 1979).

In this paper a new class of models is proposed, which, it is felt, will improve existing ones in two respects.

a. *Sensitivity of demand to accessibility and congestion is taken into account.* Most existing models assume production and/or attraction constrained spatial interactions. It is, however, sensible to assume that total demand production is not given beforehand, but is itself an endogenous variable of the spatial-interaction system. Specifically, it is reasonable to assume that demand for activities increases with accessibility. Moreover, it is sensible to assume that demand is also sensitive to congestion, that is, the more a facility is crowded, the less attractive it will be for customers. It will be shown that both features can be introduced in spatial interaction models in a very natural way, with no significant changes in their general structure.

b. *Possible combinatorial and indivisibility features in the location and size of stocks are taken into account.* Most existing models assume simple linear costs for establishing and maintaining stocks, thus being unable to account for economies of scale, bounds on the number of facilities, and bounds on their feasible sizes. On the other hand, most of these features have been introduced very well in the so-called «plant location» models developed in Operations Research (OR) (for example Erlenkotter, 1978). The usual objective function of plant location models is not very useful in Regional Science applications, since it is linear and induces an unrealistic spatial interaction behavior (i.e., users always choose the nearest destination, with no possible cross substitution). The objective functions based on Neuburger's (1971) measure of consumer surplus, however, embed realistic spatial interactions very well. It is therefore natural to take advantage of the best parts of both approaches: namely, to use a Neuburger-type objective function and plant-location-type cost functions and constraints. It will be shown how this gives rise to a family of combinatorial optimization problems of a new kind.

Although this paper will focus on the general theoretical and computational problems posed by these new models, some possible applications are discussed in Section 4. They are:

i. *The application to Lowry-type systems.* In this case the optimization concerns the location of housing and services, taking the relationships among them into account. This is not new: it has been the subject of many previous works. A completely new insight, however, is given by introducing accessibility - and congestion - sensitive mechanisms. They introduce several interesting realistic features, such as the formation of unused housing stocks, of location - dependant service - attendance ratios, and of different levels of congestion across space. Another possible improvement results from introducing combinatorial stuctures in the cost functions and in the constraints, since it is well known that real urban planning and management problems are faced with indivisibilities, thresholdlike constraints, scale effects, and the unrealistic spatial allocations obtained by continuous models.

ii. *The application to multilevel service systems.* Many services have many stages, or levels, which users will possibly go through in a given order. A typical example is given by a health care system, where users may enter the system at the lower level (usually made up of widespread small general-purpose facilities) and possibly be sent to higher, more localized levels (usually made up of larger facilities for specialized treatments). The introduction of sensitivity to accessibility and congestion is fully justified in these systems, of course, as well as indivisibilities and scale economies. Furthermore, in the example of the health care system, it is possible to have some stages in which transport between levels has the character of an emergency, rather than the normal spatial-interaction behavior. This will produce mixed models, where some stages behave according to spatial interaction, and others behave more like the OR plant location models (with possible maximum travel time constraints).

iii. *The application to multiple-destination service systems.* There are many instances where consumers make trips with many destinations, instead of home-return trips with a single destination. Most trips to retail activities are of the former kind, since usually customers have a shopping program made up of different goods, not necessarily available in the same place. But the usefulness of a round-trip scheme is not limited to shopping. Many generally different kinds of services have interactions within them, in the sense that part of the demand attracted by any one of the services may generate demand for another, depending on accessibilities. Apart from considerations similar to those for cases i and ii, it is of special interest for this case to have results on possible *aggregations*, that is, ways of building possible multipurpose facilities. This is surely relevant for retail activities, in which possible optimal patterns for shopping centers may be revealed, taking into account both spatial interaction and economies of scale.

Not all of the above problems can be solved easily, of course. Therefore, together with the general optimality conditions, approximate heuristic solution methods are developed here. Most of them are shown to be even more interesting and useful than the exact ones, since their general form is an approximate ranking rule, based on cost/benefit indicators, possibly to be improved by successive approximations. The cost/benefit indicators usually have intuitive interpretations, being made up of terms related to accessibility, congestion, demand potential, and so on.

A few general references will be given. The approach used in Section 2. to introduce accessibility and congestion sensitiveness is very closely related to the approach proposed by Walsh and Gibberd (1980), although it has been developed independently. (Earlier related work is also found in Dacey, Norcliffe, 1976; and in Jefferson, Scott, 1979). The general structure of the optimal location problem developed in Section 3. is related to the one proposed in Leonardi (1979), although it has been substantially revised and extended and findings of more recent research (Leonardi, 1980a, 1980b) have been taken into account.

2. A general accessibility - and congestion - sensitive multiactivity spatial interaction model

2.1. *A generalization of the doubly-constrained spatial interaction model*

According to the usual doubly-constrained spatial interaction model, the total number of trips for each origin-destination pair (i, j) and for a given trip purpose is determined by the set of equations

$$S_{ij} = u_i v_j f_{ij}$$

$$\sum_j S_{ij} = G_i$$

$$\sum_i S_{ij} = A_j$$

where

- S_{ij} is the number of trips from origin i to destination j
- G_i is the total number of trips generated from i
- A_j is the total number of trips attracted in j
- f_{ij} is a measure of the impedance to travel from i to j; usually, but not necessarily,
  $f_{ij} = e^{-\beta C_{ij}}$, where C_{ij} is the cost of a
  trip from i to j and β is a given nonnegative constant, called the space discount rate
- u_i, v_j are balancing factors of biproportionality.

In the models of the above type the total trip generations and attractions are usually assumed to be determined exogenously and independently of the spatial interaction process.

Let it now be assumed that the following quantities can be defined

- P_i is the maximum number of trips which can be generated from i, so that $G_i \leq P_i$; P_i will be called the *potential demand* in i
- Q_j is the maximum number of trips which can be attracted in j, so that $A_j \leq Q_j$; Q_j will be called the *total capacity* in j
- $U_i = P_i - G_i$ is the difference between the potential demand in i and the trips generated from i; U_i will be called the *unsatisfied demand* in i

$V_j = Q_j - A_j$ is the difference between the total capacity in j and the number of trips attracted in j; V_j will be called the *unused capacity* in j.

A natural assumption which can be made for endogenously determined generations and attractions is that they increase with unsatisfied demand and unused capacity, respectively. A simple mathematical translation of this assumption is given by the following equations

$$u_i = U_i/g_i \tag{1}$$

$$v_j = V_j/h_j \tag{2}$$

where g_i, h_j are given nonnegative constants. Equation (1) states that the balancing factor for the origin is proportional to the unsatisfied demand. Equation (2) states that the balancing factor for the destination is proportional to the unused capacity.

Furthermore, from the definition of U_i and V_j, the following equations hold

$$G_i + U_i = P_i \tag{3}$$

$$A_j + V_j = Q_j \ . \tag{4}$$

Substitution of (1) and (2) in the spatial interaction model and addition of (3) and (4) to the list of equations yield the following new model

$$S_{ij} = U_i V_j \frac{f_{ij}}{g_i h_j} \tag{5}$$

$$\sum_j S_{ij} = G_i \tag{6}$$

$$\sum_j S_{ij} = A_j \tag{7}$$

$$G_i + U_i = P_i \tag{8}$$

$$A_j + V_j = Q_j \ . \tag{9}$$

By means of some easy rearrangements and substitutions, model (5)-(9) can be given the following two alternative representations.

a. *The production-constrained representation*, in which the constraint on demand generation is evidenced. It is defined by the following equations

$$S_{ij} = G_i \frac{V_j f_{ij}/h_j}{\Phi_i} \qquad (10)$$

$$\Phi_i = \sum_j V_j f_{ij}/h_j \qquad (11)$$

$$G_i = P_i \frac{\Phi_i}{\Phi_i + g_i} . \qquad (12)$$

The variables Φ_i defined by (11) can be interpreted as *accessibility measures* in the Hansen sense (Hansen, 1959), and the unused capacities play the role of attractiveness measures. Equations (12) give the generated demands as functions of accessibilities. The shape of the graph of these functions is shown in fig. 1. By means of (10), (11),

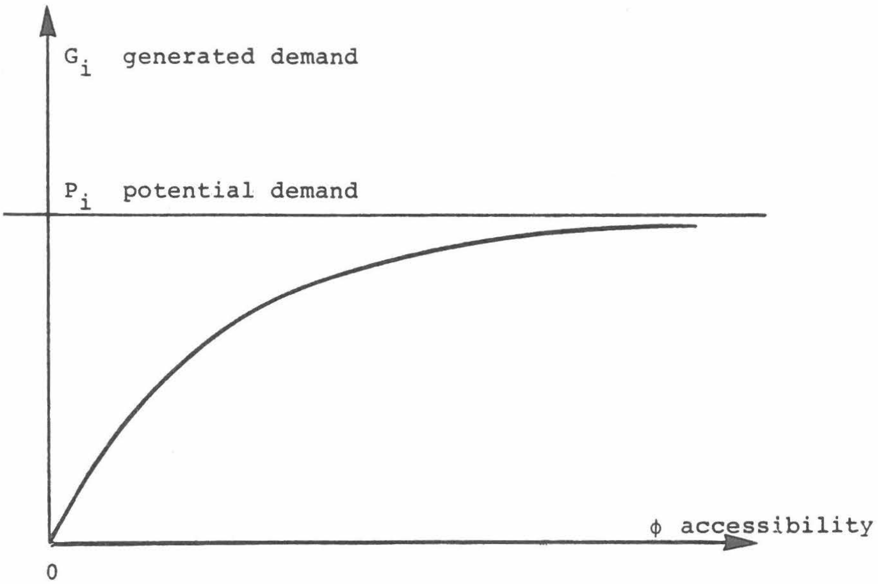


Figure 1 Generated demand as a function of accessibility

and (12) sensitiveness of generated demand both to accessibility and to congestion has been introduced. The total generated demand is a nondecreasing function of accessibility, and it tends to the total potential demand as accessibility increases. On the other hand, the accessibilities increase with the unused capacities or, which is the same, decrease as the congestion in the destinations increase.

b. *The attraction-constrained representation*, in which the constraint on demand attraction is evidenced. This representation is completely symmetrical with the production-constrained one, and is defined by the following equations

$$S_{ij} = A_j \frac{U_i f_{ij}/g_i}{\psi_j} \tag{13}$$

$$\psi_j = \sum_i U_i f_{ij}/g_i \tag{14}$$

$$A_j = Q_j \frac{\psi_j}{\psi_j + h_j} \tag{15}$$

In analogy with the interpretation given above for the Φ_i, the variables ψ_j defined by (14) can be interpreted as *population potentials* in the Stewart sense (Steward, 1948), and the unsatisfied demands represent populations. Equations (15) give the attracted demands as functions of potentials. The shape of the graph of these functions is the same as for functions (12), and is shown in fig. 2.

By means of (13), (14), and (15) sensitiveness of attracted demand both to potential and to unsatisfied demand has been introduced. The total attracted demand is a nondecreasing function of potential, and it tends to the total capacity as the potential increases. On the other hand, the potentials increase with the unsatisfied demand or, which is the same, decrease as the satisfied demand in the origins increases.

The above modified form of the classical spatial interaction model has been built by introducing the intuitive assumptions (1), (2), (3), and (4). It will be shown that this modified model can be derived by an extremal principle, which is closely related to Neuburger consumer's surplus maximization (Neuberger, 1971). Let the following function be defined

$$W(S,U,V) = -\sum_{ij} S_{ij}\left(\log \frac{S_{ij}}{f_{ij}} - 1\right) - \sum_i U_i\left(\log \frac{U_i}{g_i} - 1\right) - \sum_j V_j\left(\log \frac{V_j}{h_j} - 1\right)$$

then it can be shown that the solution to the mathematical program

$$\max_{S,U,V} W(S,U,V) \tag{16}$$

$$\text{s.t.} \sum_j S_{ij} + U_i = P_i \tag{17}$$

$$\sum_i S_{ij} + V_j = Q_i \tag{18}$$

is the spatial interaction model defined by (5)-(9). For the proof, it is first noted that the function W (S, U, V) is concave, because it is the sum of concave functions. Since constraints (17) and (18) are linear,

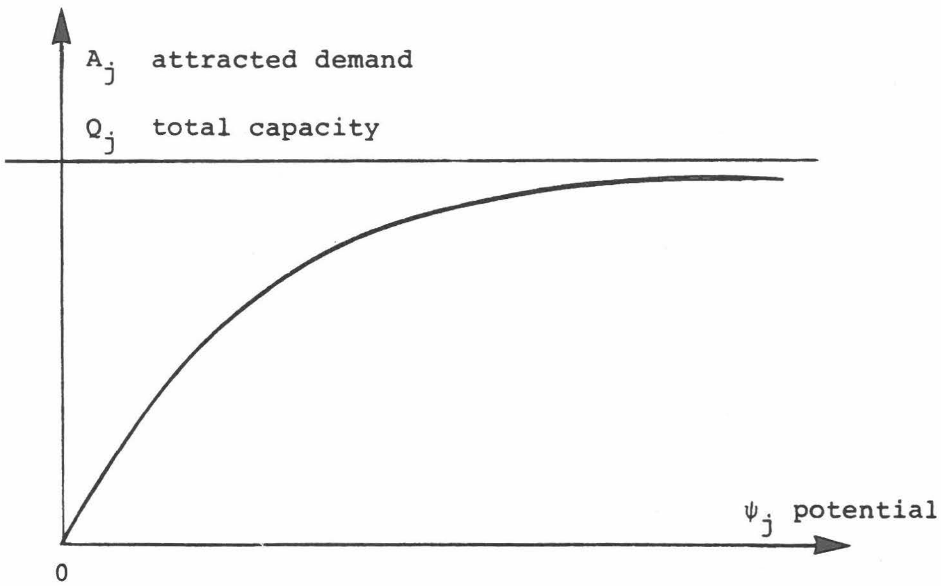


Figure 2 Attracted demand as a function of potential

(16)-(18) is a concave program, whose solution is unique. This solution must satisfy the Lagrange optimality conditions

$$\frac{\partial W}{\partial S_{ij}} - \nu_i - \mu_j = 0$$

$$\frac{\partial W}{\partial U_i} - \nu_i = 0$$

$$\frac{\partial W}{\partial V_j} - \mu_i = 0$$

or

$$-\log \frac{S_{ij}}{f_{ij}} - \nu_i - \mu_j = 0 \tag{19}$$

$$-\log \frac{U_i}{g_i} - \nu_i = 0 \tag{20}$$

$$-\log \frac{V_j}{h_j} - \mu_j = 0 \tag{21}$$

where ν_i and μ_j are the Lagrange multipliers corresponding to constraints (17) and (18), respectively. From (19), (20), and (21) it follows that

$$S_{ij} = \bar{u}_i \bar{v}_j f_{ij} \tag{22}$$

$$\bar{u}_j = U_i/g_i \tag{23}$$

$$\bar{v}_j = V_j/h_j \tag{24}$$

where

$$\bar{u}_i = e^{-\nu_i}$$

$$\bar{v}_j = e^{-\mu_j} .$$

But (23) and (24) are identical with assumptions (1) and (2), provided

$$u_i = \bar{u}_i , \quad v_j = \bar{v}_j .$$

Moreover, (17) and (18) are equivalent to (3) and (4). Hence, the solution to (16), (17), (18) satisfies equations (5)-(9).

If, as a special case, the terms depending on unsatisfied demand U_i and unused capacity V_j are dropped, then W reduces to the Neuburger consumer's surplus (except for a multiplicative constant), provided the f_{ij} are of the form

$$f_{ij} = e^{-\beta C_{ij}} .$$

2.2. *Introducing many activities*

The model discussed in Section 2.1. refers to just one trip purpose. Now let many trip purposes be introduced or, equivalently, let the trip attractors in each destination be many different activities. The following definitions will be needed.

P_i^k is the potential demand in i for activity k.

Q_j^k is the total capacity of activity k in j.

G_i^k is the demand for activity k generated in i, that is, the total number of trips from i which have an activity k as a destination.

A_j^k is the demand attracted by activity k in j, that is, the total number of trips having as a destination activity k which is located in j.

S_{ij}^k is the number of trips with purpose k (that is, having an activity k as a destination) from origin i to destination j.

f_{ij}^k is a measure of impedance to travel from i to j with purpose k; usually, but not necessarily, the measure of impedance is of the form $f_{ij}^k = e^{-\beta_k C_{ij}^k}$, where C_{ij}^k is the cost of traveling from i to j with purpose k and β_k are nonnegative space discount rates.

g_i^k, h_j^k are given nonnegative constants.

The following equations must hold

$$\sum_j S_{ij}^k = G_i^k \tag{25}$$

$$\sum_i S_{ij}^k = A_j^k \tag{26}$$

$$G_i^k + U_i^k = P_i^k \tag{27}$$

$$A_j^k + V_j^k = Q_j^k \, . \tag{28}$$

The introduction of many activities is meaningful if interactions take place among them. Let it therefore be assumed that the potential demand for each activity k from each location i, P_i^k, is not a given constant, but a linear function of the demand attracted by all activities in i. This assumption is stated by the equations

$$P_i^k = Y_i^k + \sum_r A_i^r a_{rk}$$

or, after substitution from (28)

$$P_i^k = Y_i^k + \sum_r Q_i^r a_{rk} - \sum_r V_i^r a_{rk} \tag{29}$$

where Y_i^k and a_{rk} are given nonnegative constants. The terms Y_i^k can be interpreted as exogenous inputs, while the coefficients a_{rk} are defined as follows

a_{rk} is the potential demand for activity k produced by a unit of attracted demand in activity r.

The function W(S, U, V) introduced in Section 2.1. can be generalized as follows

$$W(S,U,V) = -\sum_{ijk} S_{ij}^k \left(\log \frac{S_{ij}^k}{f_{ij}^k} - 1\right) - \sum_{ik} U_i^k \left(\log \frac{U_i^k}{g_i^k} - 1\right) - \sum_{jk} V_j^k \left(\log \frac{V_i^k}{h_j^k} - 1\right) .$$

If equations (29) are added to the list of constraints, the following generalization of (16)-(18) is obtained

$$\max_{S,U,V,P} W(S,U,V) \tag{30}$$

$$\text{s.t.} \quad \sum_j S_{ij}^k + U_i^k = P_i^k \tag{31}$$

$$\sum_i S_{ij}^k + V_j^k = Q_j^k \tag{32}$$

$$P_i^k = Y_i^k + \sum_r Q_i^r a_{rk} - \sum_r V_i^r a_{rk} . \tag{33}$$

The variables P_i^k have been added to the list of maximization variables in (30), since now they are no longer given constants. The Lagrange optimality conditions for (30)-(33) are

$$\frac{\partial W}{\partial S_{ij}^k} + \nu_i^k - \mu_j^k = 0$$

$$\frac{\partial W}{\partial U_i^k} - \nu_i^k = 0$$

$$\frac{\partial W}{\partial V_j^k} - \mu_j^k - \sum_r \lambda_j^r a_{kr} = 0$$

$$\frac{\partial W}{\partial P_i^k} + \nu_i^k - \lambda_i^k = 0$$

or

$$-\log \frac{S_{ij}^k}{f_{ij}^k} - \nu_i^k - \mu_j^k = 0 \tag{34}$$

$$-\log \frac{U_i}{g_i^k} - \nu_i^k = 0 \tag{35}$$

$$-\log \frac{V_j^k}{h_j^k} - \mu_j^k - \sum_r \lambda_j^r a_{kr} = 0 \tag{36}$$

$$v_i^k - \lambda_i^k = 0 \tag{37}$$

where $\nu_i^k, \mu_j^k, \lambda_i^k$ are the Lagrange multipliers corresponding to constraints (31), (32) and (33), respectively. From (34), (35), (36), and (37) it follows that

$$S_{ij}^k = u_i^k v_j^k f_{ij}^k \tag{38}$$

$$u_i^k = U_i^k / g_i^k \tag{39}$$

$$v_j^k = \prod_r \left(\frac{1}{u_j^r}\right)^{a_{kr}} V_j^k / h_j^k \tag{40}$$

where

$$u_i^k = e^{-\nu_i^k}$$

$$v_j^k = e^{-\mu_j^k} .$$

On the other hand, summation of (38) over j and equation (25) yield

$$G_i^k = u_i^k \Phi_i^k , \qquad \text{or} \qquad u_i^k = G_i^k / \Phi_i^k \tag{41}$$

where

$\Phi_i^k = \sum_j f_{ij}^k v_j^k$ can be interpreted as an accessibility measure, in analogy with (11).

Substitution of (41) into (40) gives for the v_j^k

$$v_j^k = \frac{V_j^k}{h_j^k} \prod_r \left(\frac{\Phi_j^r}{G_j^r}\right)^{a_{kr}} . \tag{42}$$

Equation (42) gives much insight in the way activities interact. If it is compared with its analogous (2) for the simple case, it is seen that the attractiveness of activity k in destination j is still proportional to the unused capacity V_j^k, but it is also proportional to the term

$$\prod_r \left(\frac{\Phi_j^r}{G_j^r}\right)^{a_{kr}} \tag{43}$$

that is, the product of the ratios of accessibilities to generated demands for *all activities* in j. These ratios are raised to the power a_{kr}, which is a measure of the intensity of interaction between activities k and r. Therefore, the value of (43) is mainly determined by the activities which have strong interactions with k. If, as a special case, $a_{kr} = 0$ for some r, then k and r have no interaction at all, and the corresponding factor in (43) reduces to 1. If, as a limiting case, all $a_{kr} = 0$, there is no interaction among activities, the value of (43) reduces to 1, and the value of (42) reduces to

$$v_j^k = V_j^k / h_j^k$$

which is the same as (2), except for the superscript k. In other words, the model with many activities reduces to a set of independent models with a single activity.

If (42) is substituted into Φ_i^k, the following equations are obtained

$$\Phi_i^k = \sum_j f_{ij}^k \frac{V_j^k}{h_j^k} \prod_r \left(\frac{\Phi_j^r}{G_j^r}\right)^{a_{kr}} . \tag{44}$$

In equations (44) the multiplier effect of accessibilities on themselves, and hence on demand generation, is evidenced. All accessibilities from all locations and to all activities are tied together by (44), and these ties are stronger the higher the values of the coefficients a_{kr}.

2.3. *An example: the Lowry model revisited*

The usefulness of model (30)-(33) is shown by the following example. Let an urban system be given, which is assumed to behave according to the classic economic base theory, as it has been introduced in the well known Lowry model (Lowry, 1964). In order to get a qualitative understanding for the structure of the model, some very crude simplifying assumptions will be introduced. These assumptions are the following.

a. The urban system has only two types of endogenous activities, housing and service, with no further breakdown.
b. Only one exogenous input is given, the basic sector, with no further breakdown.
c. The households are homogenous, and only the householder works.
d. The demand for housing arises only from the basic sector and the service sector.
e. The demand for services arises only from the housing sector.

Let $k = 1$ label the housing sector and $k = 2$ label the service sector, and introduce the following definitions.

S^1_{ij} is the number of households living in j, whose householders work in i.

S^2_{ij} is the number of daily trips made by households living in i to services located in j.

P^1_i is the potential demand for housing from i, that is, the total number of households whose householders work in i.

P^2_i is the potential demand for service from i, that is, the maximum number of daily trips to services which can be made by households living in i.

Q^1_j is the total capacity for housing in j, that is, the total number of dwelling units, or the housing stock, in j.

Q^2_j is the total capacity for services in j, that is, the total size of services, or the service stock, in j. (Q^2_j is assumed to be measured in terms of maximum number of daily customers that can be served).

G^1_i is the demand for housing generated in i, that is, $G^1_i = \sum_j S^1_{ij}$; in general $G^1_i < P^1_i$, that is, not all the potential demand for housing is necessarily satisfied.

G^2_i is the demand for service generated in i, that is, $G^2_i = \sum_j S^2_{ij}$; in general $G^2_i < P^2_i$, that is, the maximum number of possible trips to services is not necessarily made.

A^1_j is the demand for housing attracted in j, that is, $A^1_j = \sum_i S^1_{ij}$; in general $A^1_j < Q^1_j$, that is, not all the housing stock is necessarily used.

A_j^2 is the demand for service attracted in j, that is, $A_j^2 = \sum_i S_{ij}^2$; in general $A_j^2 < Q_j^2$, that is, not all the service capacity is necessarily used.

Y_i^1 is the number of households whose householder works in the basic sector.

a_{12} is the potential number of daily trips to services made by a household.

a_{21} is the ratio between workers in the service sector and total attracted service demand.

Assumptions b, d, and e imply that $Y_i^1 = 0$, $a_{11} = 0$, $a_{22} = 0$. Equations (29) assume the simple form

$$P_i^1 = Y_i^1 + A_i^2 a_{21} \tag{45}$$

$$P_i^2 = A_i^1 a_{12} \ . \tag{46}$$

Equation (45) states that the total potential demand for housing is equal to the number of workers in the basic sector, plus the number of workers in the service sector. Equation (46) states that the total potential demand for service is equal to the maximum number of daily trips to services which can be made by households.

If equation (42) is applied to the housing sector, it takes the form

$$v_j^1 = \frac{V_j^1}{h_j^1}\left(\frac{\Phi_j^2}{G_j^2}\right)^{a_{11}} \tag{47}$$

where

$V_j^1 = Q_j^1 - A_j^1$ is the unused housing stock in j

Φ_j^2 is the accessibility to services from j.

Therefore, the attractiveness of location j as a place of residence, as measured by v_j^1, increases both with the availability of dwelling units, V_j^1, and with the accessibility to services, Φ_j^2. The main trade-off in residential choice is thus embodied in (47). The third trade-off term, the home-to-work travel cost, is introduced if the production-constrained representation (see Section 2.1.) is used for the S_{ij}^1

$$S_{ij}^1 = G_i^1 \frac{V_j^1\left(\frac{\Phi_j^2}{G_j^2}\right)^{a_{12}} f_{ij}^1/h_j^1}{\Phi_i^1} \tag{48}$$

where $\Phi_i^1 = \sum_j v_j^1 f_{ij}^1$, and v_j^1 is defined as in (47). Since f_{ij}^1 dependes on the cost of traveling from i to j, plus possibly some additional costs associated with location j (like the rent), (48) shows how availability of houses, accessibility to services, home-to-work travel cost and location costs determine the overall attractiveness for residential location. Other results easily derived from the production-constrained representation are

$$G_i^1 = P_i^1 \frac{\Phi_i^1}{\Phi_i^1 + g_i^1} \qquad \text{the demand for housing generated in j} \tag{49}$$

$$U_i = P_i^1 \frac{g_i^1}{\Phi_i^1 + g_i^1} \qquad \text{the unsatisfied demand for housing in i} \tag{50}$$

$$A_j^1 = \frac{V_j^1}{h_j^1} \left(\frac{\Phi_j^2}{G_j^2}\right)^{a_{12}} \sum_i \frac{G_i^1}{\Phi_i^1} f_{ij}^1 \qquad \text{the housing demand attracted in j.} \tag{51}$$

Equation (51) can be given a more meaningful and simpler form. From (49) and (50) it follows that

$$U_i^1 = \frac{G_i^1}{\Phi_i^1} g_i^1$$

therefore

$$\sum_i \frac{G_i^1}{\Phi_i^1} f_{ij}^1 = \sum_i U_i^1 f_{ij}^1 / g_i^1 = \psi_j^1$$

is the unsatisfied housing demand potential, as defined in the attraction-constrained representation of Section 2.1.

Substitution of this result into (51) yields

$$A_j^1 = \frac{V_j^1}{h_j^1} \left(\frac{\Phi_j^2}{G_j^2}\right)^{a_{12}} \psi_j^1 \ . \tag{52}$$

Equation (52) embodies the spatial interaction process in the most synthetic and intuitive way. It says that the total housing demand attracted in j is proportional to the availability of houses in j, V_j^1, to the potential ψ_j^1, which is a measure of nearness of j to unsatisfied housing demand, and to the accessibility to services from j, raised to the power a_{12}, which is the maximum daily frequency of home-to-service trips. From (52) an equation for the unused housing stock is easily derived. If:

$$A_j^1 = Q_j^1 - V_j^1$$

is substituted for A_j^1 in (52), and the resulting equations is solved for V_j^1, it is found that

$$V_j^1 = Q_j^1 \frac{h_j^1}{\left(\frac{\Phi_j^2}{G_j^2}\right) \psi_j^1 + h_j^1} . \tag{53}$$

Equation (53) says that the formation of unused housing stock mainly takes place in locations far from both services and from places of work, where the housing demand arises. This is exactly what might be expected. However, from (52) it is seen that the unused housing stock is an attracting factor for new housing demand. Therefore, the housing demand is forced to trade off accessibility to services and nearness to the place of work (which would solely guide their choice, other things being equal) with availability of houses, which acts as a constraint. The resulting spatial pattern is a concentration of households in locations with highest accessibility to services and places of work, whose housing capacity is near to saturation, and a lower density of households in the less accessible locations, where unused housing stock may possibly be found. This is indeed very close to what actually happens in real urban systems, and it is also very similar to what the classic Lowry model predicts. However, when total demand grows faster than the housing stock, all locations tend to be saturated, whether their accessibility is high or low. This behavior is also very close to the actual behavior of congested urban systems, but it cannot be accounted for by the classic «unconstrained» Lowry model.

The analysis which has been carried out for the housing sector applies to the service sector as well, and it will not be repeated here.

3. The optimal location problem

3.1. *The primal problem*

In Section 2. the analysis of the descriptive process has been carried out. Now the problem of how to control the multiactivity spatial interaction system in some optimal way will be posed. That is, given that customers behave as if they were looking for the optimal solutions to problems (30)-(33), how can a public authority improve this optimizing behavior by suitably choosing the values for the physical stocks of activities, that is, the capacitites Q_j^k? The above question implies the assumption that the goal of the customers (maximizing the function W defined in Section 2.2.) is in agreement with that of the public authority, so that no conflicting-goal problem arises between customers and public authority. The public authority is also assumed to pay the costs to establish the capacities Q_j^k. Let the cost functions be of the form

$$a_j^k + b_j^k Q_j^k$$

where a_j^k is a fixed-charge cost to be paid for establishing an activity k in location j, while b_j^k is a unit cost. Fixed charges have the effect of introducing economies of scale and threshold effects, as it will be shown later.

The optimization problem can be split in two steps:

a. choose a subset of locations and a subset of activities to be established for each chosen location;

b. given the result of step a, find the optimal size of the activities to be established in each chosen location.

While step a gives rise to a combinatorial problem, step b is a smooth mathematical programming problem. Let therefore step b be solved first, and step a be introduced in the next section. It will be thus assumed that the chosen locations and activities are given, and only the capacities Q_j^k have to be found. The resulting mathematical programming problem is the same as (30)-(33), the only difference being that establishing costs are subtracted from the objective function W, and the capacities Q_j^k are added to the list of decision variables:

$$\max_{S,U,V,P,Q} \ W(S,U,V) - \lambda \left(\sum_{jk} a_j^k + \sum_{jk} Q_j^k b_j^k\right) \tag{54}$$

$$\text{s.t.} \quad \sum_j S_{ij}^k + U_i^k = P_i^k \tag{55}$$

$$\sum_i S_{ij}^k + V_j^k = Q_j^k \tag{56}$$

$$P_i^k = Y_i^k + \sum_r Q_i^r a_{rk} - \sum_r V_i^r a_{rk} \ . \tag{57}$$

The parameter λ which multiplies the cost term in (54) is a trade-off parameter, weighting costs against benefits. Usually a sensitivity analysis has to be made on λ, in order to assess the appropriate trade-off level.

Alternatively, λ may be interpreted as a Lagrange multiplier arising from the relaxation of a budget constraint.

Formulation (54)-(57) is somewhat redundant. First, since the locations and activities are assumed to be given, the sum of the fixed-charge costs is constant, and can be dropped from the objective function (54). Secondly, there is no use to keep P_i^k and Q_i^k in the list of decision variables, since by means of equations (55) and (56) they can be expressed in terms of the variables S_{ij}^k, U_i^k, V_j^k. Therefore, after some substitutions and rearrangements, problem (54)-(57) reduces to:

$$\max_{S,U,V} W(S,U,V) - \lambda \sum_{jk} b_j^k (\sum_i S_{ij}^k + V_j^k) \tag{58}$$

$$\text{s.t.} \quad \sum_j S_{ij}^k + U_i^k - \sum_r a_{rk} \sum_j S_{ji}^r = Y_i^k \,. \tag{59}$$

Problem (58)-(59) will be referred to as the «primal» problem.

3.2. *Some duality results*

The saddle-point problem equivalent to (58)-(59) is

$$\min_{\nu} \max_{S,U,V} L(S,U,V,\nu)$$

where the Lagrangean function L is defined as:

$$L(S,U,V,\nu) = W(S,U,V) - \lambda \sum_{jk} b_j^k (\sum_i S_{ij}^k + V_j^k) + \\ + \sum_{jk} \nu_i^k (Y_i^k - \sum_j S_{ij}^k - U_i^k + \sum_r a_{rk} \sum_j S_{ji}^r)$$

and the ν_i^k are the Lagrange multipliers, or dual variables, corresponding to constraints (59). The vanishing of the derivatives of L with respect to the primal variables yields the following equations

$$\left.\begin{aligned} -\log \frac{S_{ij}^k}{f_{ij}^k} - \lambda b_j^k - \nu_i^k + \sum_r a_{kr} \nu_j^r = 0 \,, \quad \text{or} \\ S_{ij}^k = f_{ij}^k e^{-(\lambda b_j^k + \nu_i^k - \sum_r a_{kr} \nu_j^r)} \end{aligned}\right\} \tag{60}$$

$$-\log \frac{U_i^k}{g_i^k} - \nu_i^k = 0 \,, \quad \text{or} \quad U_i^k = g_i^k e^{-\nu_i^k} \tag{61}$$

$$- \log \frac{V_j^k}{h_i^k} - \lambda b_j^k = 0 \ , \qquad \text{or} \qquad V_j^k = h_j^k e^{-\lambda b_j^k} \ . \tag{62}$$

By means of equations (60), (61), and (62) the primal variables can be expressed in terms of the dual variables in closed form. Substitution into *L* and some rearrangements yield the following unconstrained «dual» problem

$$\min_{\nu} D(\nu)$$

where the dual objective function D is defined as

$$D(\nu) = \sum_{ijk} S_{ij}^k(\nu) + \sum_{jk} U_i^k(\nu) + \sum_{jk} V_j^k + \sum_{jk} \nu_i^k Y_i^k \tag{63}$$

and the functions $S_{ij}^k(\nu)$, $U_i^k(\nu)$, and the constants V_j^k are defined by equations (60), (61), and (62). The dual objective function can be given an intuitive interpretation. From equation (32), the total capacity of activity k in location j, Q_j^k, is given by

$$Q_j^k = \sum_i S_{ij}^k + V_j^k$$

so that

$$Q_j^k(\nu) = \sum_i S_{ij}^k(\nu) + V_j^k \tag{64}$$

is the total capacity of activity k in location j, as a function of the dual variables.

If (64) is substituted into (63), the dual function becomes

$$D(\nu) = \sum_{jk} U_i^k(\nu) + \sum_{jk} Q_j^k(\nu) + \sum_{jk} \nu_i^k Y_i^k \ . \tag{65}$$

The first two terms of (65) are the *total unsatisfied demand* and the *total capacity*, respectively. The philosophy behind minimization of D(ν) is therefore a balance between a welfare goal (minimizing unsatisfied demand) and an efficiency goal (minimizing the total capacity). Now let the combinatorial part of the problem (step a of Section 3.1.) be introduced. Define the boolean variables

$$x_j^k = \begin{cases} 1, & \text{if an activity k is located in j} \\ 0, & \text{otherwise.} \end{cases}$$

Further constraints may be introduced on the number of activities which can be established in the same location. For simplicity, it will be provisionally assumed that only one activity can be established in each location. The assumption seems restrictive, but it may be easily relaxed, as it will be done in later sections. In terms of the boolean variables, the assumption gives rise to the constraints

$$\sum_k x_j^k \leq 1 \ .$$

If the boolean variables and the sum of the fixed charges (which now is no longer a constant) are suitably introduced in (65), the following modified dual function is obtained

$$D(\nu,x) = \sum_{jk} [U_i^k(\nu) + \nu_i^k Y_i^k] + \sum_{jk} x_j^k [Q_j^k(\nu) - \lambda a_j^k] \ . \tag{66}$$

This function has to be minimized with respect to the dual variables ν_i^k and maximized with respect to the variables x_j^k. The resulting problem is

$$\max_x \min_\nu D(\nu,x) \tag{67}$$

$$\text{s.t.} \quad \sum_k x_j^k \leq 1 \tag{68}$$

$$x_j^k \in \{0,1\} \ . \tag{69}$$

An upper bound to the optimal value of D is obtained by relaxing constraint (69) and replacing it with the weaker condition:

$$0 \leq x_j^k \leq 1 \tag{70}$$

where the variables x_j^k are allowed to assume any real value in the unit interval. But the right-hand side inequality in (70) is redundant, since it is already implied by constraints (68). Therefore, the relaxed version of (67)-(69) becomes

$$\max_x \min_\nu D(\nu,x) \tag{71}$$

$$\text{s.t.} \quad \sum_k x_j^k \leq 1 \tag{72}$$

$$x_j^k \geq 0 \ . \tag{73}$$

Problem (71)-(73) is a saddle-point problem. It is therefore natural to look at $D(\nu, x)$ as the Lagrangean function associated with some «primal» problem, the variables x_j^k playing the role of Lagrange multipliers. It will be shown that such a «primal» problem indeed exists, and it is given by the following mathematical program

$$\min_{\nu,z} P(\nu,z) \tag{74}$$

$$\text{s.t.} \quad z_j \geq Q_j^k(\nu) - \lambda a_j^k \tag{75}$$

$$z_j \geq 0 \tag{76}$$

where the function $P(\nu,z)$ is defined as

$$P(\nu,z) = \sum_{jk} [U_i^k(\nu) + \nu_i^k Y_i^k] + \sum_j z_j \ .$$

To show that (74)-(76) is equivalent to (71)-(73), the following «Lagrangean» function is introduced

$$\bar{D}(\nu,z,x,\varepsilon) = P(\nu,z) - \sum_{jk} x_j^k [z_j - Q_j^k(\nu) - \lambda a_j^k] - \sum_j \varepsilon_j z_j$$

where x_j^k and ε_j are the Lagrange multipliers corresponding to constraints (75) and (76), respectively. Problem (74)-(76) is equivalent to the following saddle-point problem

$$\max_{x,\varepsilon} \min_{\nu,z} \bar{D}(\nu,z,x,\varepsilon) \tag{77}$$

$$\text{s.t.} \quad x_j^k \geq 0 \tag{78}$$

$$\varepsilon_j \geq 0 \ . \tag{79}$$

(The nonnegativity constraints on the multipliers are required because constraints (75) and (76) are inequalities). The vanishing of the derivatives of $\bar{D}$ with respect to z_j implies

$$1 - \sum_k x_j^k - \varepsilon_j = 0$$

or

$$\sum_k x_j^k = 1 - \varepsilon_j \ . \tag{80}$$

Equation (80) and constraints (78) and (79) imply

$$0 \leq 1 - \varepsilon_j \leq 1$$

therefore (80) is equivalent to (72). Substitution of (80) into (77) yields:

$$\begin{aligned}\bar{D}(\nu, z, x, \varepsilon) = & \sum_{jk} [U_i^k(\nu) + \nu_i^k Y_i^k] + \sum_j z_j - \\ & - \sum_j (1 - \varepsilon_j) z_j + \sum_{jk} x_j^k [Q_j^k(\nu) + \lambda a_j^k] - \sum_j \varepsilon_j z_j\end{aligned}$$

and since the terms in z_j and ε_j cancel out a comparison with (66) shows that

$$\bar{D}(\nu, z, x, \varepsilon) = D(\nu, x) \ .$$

It follows that problem (77)-(79) is equivalent to problem (71)-(73), and hence that problem (74)-(76) is equivalent to problem (71)-(73).

A more detailed description of the general structure of the solution to (74)-(76) will now be given. The way the function $P(\nu, z)$ has been built always forces the variables z_j to assume the lowest possible value in the optimal solution of (74)-(76). From constraints (75) and (76) it follows that it must be either

$$z_j = \max_k [Q_j^k(\nu) - \lambda a_j^k] \tag{81}$$

or

$$z_j = 0 \tag{82}$$

or both, which ever is greater. When only (81) holds, a k* which maximizes its right-hand side exists such that

$$z_j = Q_j^{k*}(\nu) - \lambda a_j^{k*} > 0 \tag{83}$$

while for every $k \neq k^*$ it must be

$$z_j > Q_j^k(\nu) - \lambda a_j^k \ . \tag{84}$$

Therefore, constraint (75) is binding for activity k* only, and nonbinding for all other activities. It follows that the multipliers of constraints (75) are

$$x_j^{k^*} \geq 0 \tag{85}$$

$$x_j^k = 0 \qquad \text{for} \qquad k \neq k^* \ . \tag{86}$$

On the other hand, since (82) does not hold, constraint (76) is nonbinding, therefore it must be

$$\varepsilon_j = 0 \ . \tag{87}$$

Substitution of (85), (86) and (87) into (80) yields

$$x_j^{k^*} = 1 \tag{88}$$

that is, the location j is chosen to establish an activity k*. It is important to notice that (88) yields a natural integer solution, that is, one which is feasible for the original combinatorial problem.

When only (82) holds, then it follows that

$$Q_j^k(\nu) - \lambda a_j^k < 0 \qquad \text{for all } k \tag{89}$$

that is, constraints (75) are nonbinding for all activities, and the corresponding multipliers will be

$$x_j^k = 0 \qquad\qquad \text{for all } k \ . \tag{90}$$

In other words, no activity will be established in location j. Again it is important to notice that (90) yields a natural integer solution.

When both (81) and (82) hold, then (85) and (86) hold as well, but instead of (87) it must be

$$\varepsilon_j \geq 0$$

and from (80) it follows that

$$x_j^{k^*} = 1 - \varepsilon_j \leq 1 \ . \tag{91}$$

That is, a natural integer solution is no longer assured and fractional values for $x_j^{k^*}$ may be (and usually are) introduced.

Equations (81) and (82) suggest a new possible formulation of problem (74)-(76). From (81) and (82) it follows that

$$z_j \doteq \max \{ \max_k [Q_j^k(v) - \lambda a_j^k] \ , \ 0 \}$$

and if this result is substituted in $P(v,z)$ the following non-smooth optimization problem is obtained:

$$\min_v G(v) \tag{92}$$

where the function $G(v)$ is defined as

$$G(v) = \sum_{jk} [U_i^k(v) + v_i^k Y_i^k] + \sum_j \max \ \{ \max_k [Q_j^k(v) - \lambda a_j^k], \ 0 \} \ .$$

Problem (92) is computationally attractive, since it is unconstrained and contains only the variables v_i^k. The price to be paid for this simplicity is the nonsmoothness of the function G.

A summary of the main duality and equivalence results is useful. If $\bar{x}$, $\bar{v}$, $\bar{z}$ denote the optimal values for the corresponding arrays of variables, the following equalities hold

$$D(\bar{v}, \bar{x}) = G(\bar{v}) = P(\bar{v}, \bar{z}).$$

For general nonoptimal values x, v, z the following inequalities hold

$$D(v, x) \leq G(v) \leq P(v, z). \tag{93}$$

If, as a special case, x is the optimal integer solution, it is seen from (85) that $G(v)$ provides the tighter upper bound to $D(v, x)$. Anyway, both G and P can be used to compute upper bounds to D, depending on computational convenience. Problem (92) is simple, but nonsmooth, as already stated. Problem (74)-(76) is a smooth convex programming problem but has the nonlinear constraints (75). If an algorithm to solve either (92) or (74)-(76) is available, it can be used to find the optimal relaxed values for the x_j^k. If all the x_j^k assume natural integer values, then the original combinatorial problem is solved, and no further refinement is needed. If some x_j^k assume fractional values, a branch-bound refinement procedure may be started.

3.3. *Heuristic approximations*

The problems introduced in Section 3.2. may be hard to solve exactly, and the requirement of integer values for the variables x_j^k makes the task even harder.

However, there are many reasons why applications to real problems should not be obsessed with finding exact solutions. First, the input data and the definition of the physical setting are always less precise than an exact algorithm seems to imply. The set of possible locations, for instance, is usually a set of zones in which a given area is subdivided, and there is much arbitrariness in this subdivision. An exact algorithm would possibly be very sensitive to changes in the subdivision, but in the real world such changes are meaningless. Secondly, finding an exact solution corresponding to a given set of input data is much less interesting and useful than having a whole spectrum of solutions corresponding to different sets of input data. A sensitivity analysis has typically to be carried out on parameters like the space discount rate, the travel costs, the trade-off between benefits and costs, the elasticity of demand to accessibility, the minimum feasible size for the activities, and so on. Finding an exact solution for all the possible combinations of different values for these parameters is usually prohibitive. Third, producing numerical solutions is not the only aim of optimal location models, nor is it necessarily the main one. Qualitative understanding of the relationships among the main factors affecting location patterns is often a much more interesting goal, both in theory and in applications.

The reasons listed above suggest that fast and easy heuristic approximations could be a useful tool for optimal location problems.

A heuristic approach to solving (74)-(76), subject to the integrality conditions on the multipliers x_j^k, may be developed starting from equations (81)-(91), which can be summarized as follows

let k* maximize

$$Q_j^k(\nu) - \lambda a_j^k$$

then

$$\text{if } Q_j^{k*}(\nu) - \lambda a_j^{k*} > 0 \,, \qquad x_j^{k*} = 1 \tag{94}$$

$$\text{if } Q_j^{k*}(\nu) - \lambda a_j^{k*} = 0 \,, \qquad 0 \le x_j^{k*} \le 1 \tag{95}$$

$$\text{if } Q_j^{k*}(\nu) - \lambda a_j^{k*} < 0 \,, \qquad x_j^{k*} = 0 \,. \tag{96}$$

The above result refers to general (noninteger) values for x_j^k. Since, however, an integer solution is looked for, it will be assumed that only integer values will be introduced in the trial solutions. Therefore equation (95) can be dropped and the following heuristic optimality conditions are obtained:

$$\text{if} \quad Q_j^{k*}(v) \geq \lambda a_j^{k*} , \qquad x_j^{k*} = 1 \tag{97}$$

$$\text{if} \quad Q_j^{k*}(v) < \lambda a_j^{k*} , \qquad x_j^{k*} = 0 . \tag{98}$$

Conditions (97) and (98) state a very reasonable efficiency principle. If the total size required by the activity k* in location j is greater than, or equal to, the fixed charge a_j^{k*} to be paid for establishing it (multiplied by the trade-off parameter λ), than j is a good location for activity k*, and it is worth establishing it there. If the total size required by the activity k* in location j is less than the fixed charge term, then j is a bad location for activity k*, which will not be established there. From (97) and (98) a very simple interpretation of the fixed-charge term follows: λa_j^k *is the minimum feasible size for an activity k in location j.* This interpretation is very useful in applications, since it is often easier to assess the values for the minimum feasible sizes, rather than for the fixed costs. Conditions (97) and (98) can be rephrased in the following first rule of thumb (ROT1):

ROT1 *choose only those locations where at least one activity requires a capacity at least as great as the minimum feasible size; establish in each of these locations only the activity with the highest difference between required capacity and minimum size.*

The rational behind ROT1 is that only those activities will be established that attract enough demand to justify at least the minimum feasible size. The reason why only one activity is possibly established in each location is because of contraints (72). But now these constraints can be easily relaxed, and the more general case, in which many different activities can be established in the same location, can be introduced. The efficiency conditions for this case are stated in the following, second rule of thumb (ROT2):

ROT2 *step 1* *for each possible location, rank the activities according to the difference between required capacity and minimum feasible size; if this difference is negative, drop the corresponding activities from the list*

step 2 *establish in each location as many activities as possible, choosing them according to the ranking obtained in step 1.*

There is some vagueness in step 2, since the precise meaning of «as many activities as possible» has not been defined. However, it is felt that it is better to keep this vagueness, and leave the decision maker some freedom on judging by inspection when to stop picking up activities from the list. This freedom is needed because the constraints imposed in each zone by the limited availability of space and by the already existing physical stocks act in a nonsmooth and hardly quantifiable way. The tools provided by ROT2 do not solve the problem of how to meet these constraints in the best way, which is left to town designers and architects. ROT2 simply yields a set of indicators by means of which activities and locations can be ranked for a possible choice.

It is worth recalling that, although the indicators and the ranking produced by ROT2 are very simple and intuitive, they are rooted in a rigorous ground, since ROT2 has been obtained by suitably approximating and generalizing the exact optimality conditions (94), (95), and (96).

Both ROT1 and ROT2 can be used to generate improved values for the variables x_j^k. The general structure of a possible iterative algorithm is shown in the block diagram of fig. 3.

The diagram is self-explanatory and only a few comments are needed. The function $D(v, x)$ is the one defined by (66). Step 1 is quite arbitrary and any initial guess can be used. However, since the algorithm is a heuristic one, independence of the final solution from the initial guess may not be assured. Therefore, it is worth putting some effort in finding a good initial guess. When no better assumption is available, two possible starts are:

1. all activities are established in all locations;
2. no activity is established in any location.

Although assumption 1 may seem more reasonable, assumption 2 has some definite computational advantages. This can be shown by performing the first iteration of the algorithm. If all $x_j^k = 0$, then

$$\min_v D(v, x)$$

reduces to

$$\min_v \sum_{jk} (g_i^k e^{-v_i^k} + v_i^k Y_i^k) \tag{99}$$

[(99) follows from (61) and (66)]. Standard calculus yields the solution to (99)

$$e^{-v_i^k} = \frac{Y_i^k}{g_i^k}, \qquad \text{or} \qquad e^{v_i^k} = \frac{g_i^k}{Y_i^k} \tag{100}$$

and substitution of (100) into (64) yields the total capacities required in each location

$$Q_j^k = e^{-\lambda b_j^k}\left[\sum_i f_{ij}^k \frac{Y_i^k}{g_i^k} \prod_r \left(\frac{g_j^r}{Y_j^r}\right)^{a_{kr}} + h_j^k\right] \tag{101}$$

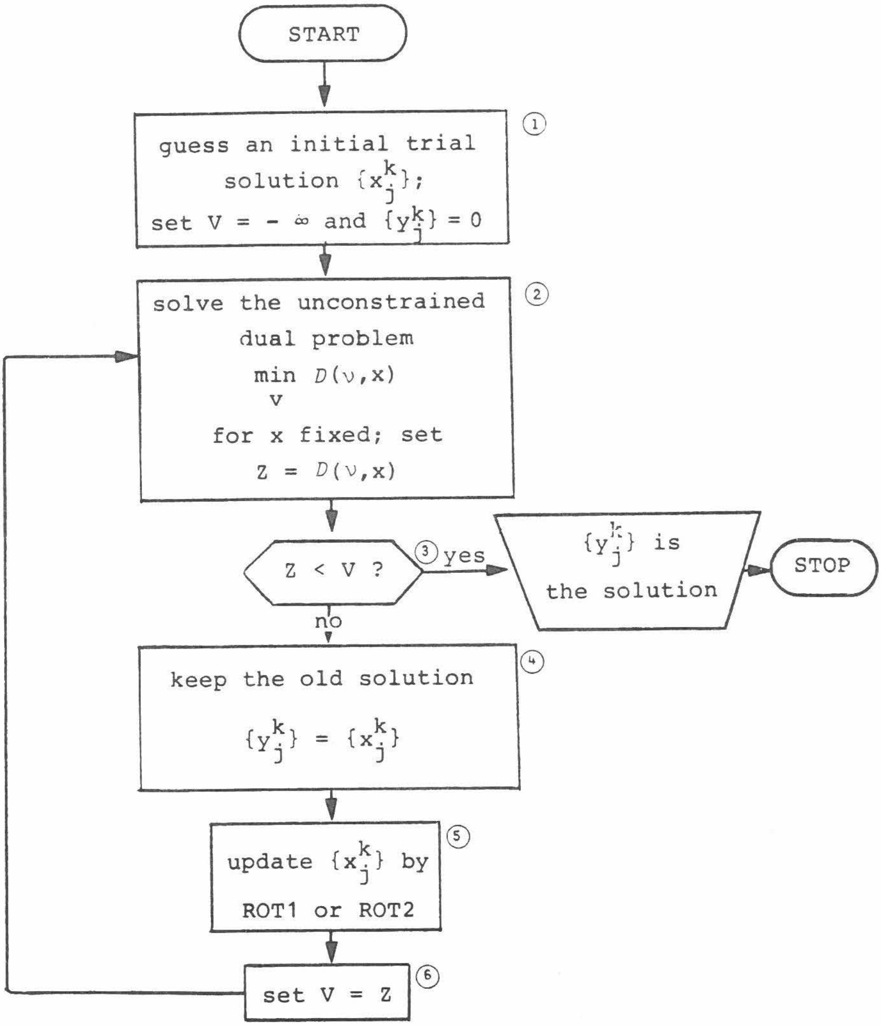


Figure 3 An iterative heuristic algorithm for the multiactivity location problem

The right-hand side of (101) depends only on given constant terms, and can be computed beforehand. Going now to step 5 of the algorithm, the differences

$$Q_j^k - \lambda a_j^k \tag{102}$$

are computed and used to update the x_j^k, either by ROT1 or by ROT2. A sensible updating can be performed only if some of the differences (102) are positive. If, however, it is found at this step that *all* the differences (102) are negative, the algorithm stops after two iterations, and the final solution is the same as the starting guess 2, that is doing nothing. When this happens, no location problem exists. It is therefore suggested to use the start 2, in order to let the algorithm check for this possibility from the very beginning. It is also suggested that, when the algorithm has such a stop, input data should be carefully checked for possible mistakes.

The method suggested for step 2 is computationally the best one, since the mathematical program

$$\min_{\nu} D(\nu, x)$$

for x fixed reduces to

$$\min_{\nu} D(\nu) \tag{103}$$

where $D(\nu)$ is the function defined by (63) or (65). Problem (103) is a simple convex unconstrained minimization problem, which can be solved by standard techniques. However, it might be felt that it is rather abstract, since it is defined in terms of the dual variables, to which no immediate intuitive meaning can be given. It may therefore be useful to reformulate the resulting capacities Q_j^k in terms of the more «physical» quantities and indicators introduced in Section 2.2. Some rearrangements of (32), (38), (39), and (42) yield

$$Q_j^k = V_j^k \left[\frac{1}{h_j^k} \prod_r \left(\frac{\Phi_j^r}{G_j^r} \right)^{a_{kr}} \psi_j^k + 1 \right] \tag{104}$$

where all the variables have been already defined in Sections 2.1. and 2.2. The main definitions will be briefly restated

V_j^k is the unused capacity of activity k in location j

Φ_j^r is the accessibility to activities r from location j

ψ_j^k is the potential of unsatisfied demand for activity k in location j

G_j^r is the demand for activity r generated in location j.

Equation (104) gives an intuitive interpretation of the mechanism implied by (103), and indicators like accessibilities and potentials give further elements to evaluate the resulting location pattern. It might be argued that some iterative scheme based on (104), together with (31), (32), (33), could be devised, without resorting to the dual formulation (103). Such a scheme could actually be built, in close analogy with the original method proposed by Lowry (1964). However, it would be computationally very poor, compared to the efficiency of (103).

In step of fig. 3 a stopping rule based on the value of the dual objective function is suggested. A stronger stopping rule could be based on the array $\{x_j^k\}$, that is: *stop when the same* $\{x_j^k\}$ *is obtained in two successive steps.* However, such a condition may possibly never be met, since the x_j^k can assume only integer values. Cycling may occur, indicating that possible multiple solutions exist, or that the algorithm is not able to give any further improvement. The rule based on the value of the dual objective function seems therefore more stable.

The algorithm of fig. 3 assumes that all the parameters are held constant. It will now be shown how it can be generalized to perform some sensitivity analysis. As an example, let it be assumed that a sensitivity analysis on the trade-off parameter λ is needed. A possible algorithm is shown in the block diagram of fig. 4. In this algorithm it has been assumed that only ROT2 is used. A few explanations will be given. The functions $\mathrm{D}(v)$ and $D(v, \mathrm{x})$ are defined by equations (65) and (66), respectively. The sensitivity analysis starts with $\lambda = 0$, that is no costs are paid to establish the activities (alternatively, no constraint is put on the minimum feasible size of the activities). $D(v, \mathrm{x})$ reduces $\mathrm{D}(v)$, and all activities are established in all locations(*). This is the meaning of the initial steps 1 and 2. In step 3 a nonzero trade-off parameter is introduced, and its initial value is set equal to a given step size τ. Steps 5-11 closely resemble the routine 2-6 of the algorithm of fig. 3. There are two major differences, however. First, the initial trial solution is replaced by the solution produced in the last iteration over λ. It is argued that small changes in λ produce small changes in the array $\{x_j^k\}$, so that the optimal $\{x_j^k\}$ for a given λ should be a good start to find the optimal $\{x_j^k\}$ for $\lambda + \tau$. In this way, the information gained at each iteration over λ is used to speed up the convergence at the next iteration. Secondly, the new stopping rule 7 has been introduced. Its meaning has already been discussed in

(*) Here it is assumed for simplicity that ROT2 is used with no constraints on the number of activities to be established in each location. A modified version of the algorithm which takes such constraints into account can be easily developed.

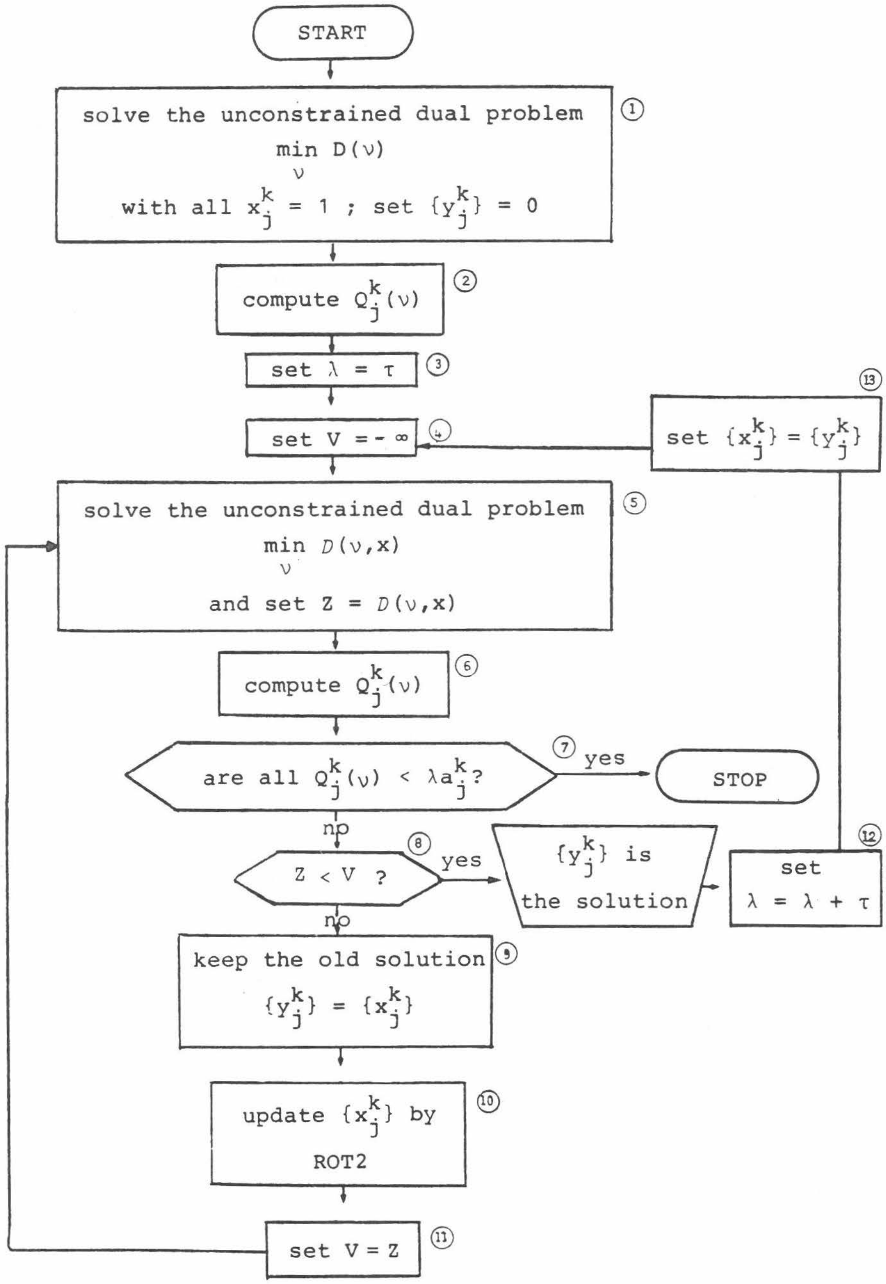


Figure 4 An iterative heuristic algorithm for the multiactivity location problem with sensitivity analysis on the trade-off parameter

connection with the problem of the choice of an initial start for the algorithm of fig. 3. When the value of λ is such that

$$Q_j^k(\nu) < \lambda a_j^k \qquad \text{for all j and k} \tag{105}$$

then no location problem exists, and no further sensitivity analysis over λ is needed. However, it should be noticed that the logic behind the algorithm of fig. 3 here is turned upsidedown. For the algorithm of fig. 3 a start with no activity in no location has been suggested. In the algorithm of fig. 4 a start with all activities in all locations is required, and conditions (105) are used as a *final* stopping rule. This is because in the algorithm of fig. 4 conditions (105) are reached in a natural way, and have nothing to do with possible mistakes or inconsistencies in the input data.

In step 12 of fig. 4 the value of λ is updated by incrementing it with the step size τ, and the whole process is repeated.

The above approach can be extended to include the sensitivity analysis on other meaningful parameters, like the space discount factor. The details of these extensions will not be developed here.

4. Some applications

4.1. *The urban system*

The descriptive model of a simplified urban system has been already outlined in Section 2.3. Here it will be shown how this model can be embedded in a mathematical program which optimizes the location of housing and services. The same notation and assumption of Section 2.3. will be kept, and the following further assumptions will be introduced.

a. The unsatisfied demand for housing is always zero, that is, the housing demand is accessibility insensitive. This assumption implies that the housing stock will always be made big enough to satisfy all the demand. It also implies

   $g_i^1 = 0 \qquad$ for all i.

b. The unused capacity for services is always zero, that is, the service demand is congestion insensitive. This assumption implies

   $h_j^2 = 0 \qquad$ for all j.

c. The costs for establishing the housing and the service facilities consists of the fixed charge term only. Otherwise stated, instead of establishing costs, minimum size requirements are introduced. Moreover, the minimum feasible sizes are the same for all locations.

d. No constraints are placed on the number of different activities which may be established in each location, that is, housing and services may always be located in the same place.

The above assumptions have been introduced for sake of simplification, although there is no real computational obstacle to solve the problem in its more general form.

Assumption a prevents the formation of unsatisfied housing demand. The existence of such a demand implies introducing phenomena like cohabitation, overcrowding, and formation of slums, which are surely realistic features of many real urban systems. However, it is felt that such phenomena need the introduction of variables other than accessibility and available capacity to be fully explained, and this would go beyond the scope of this simplified example. Assumptions b and c imply that all the service facilities are equally attractive, and only their accessibility determines the customers' choice. This is clearly an oversimplification in the more realistic case where the services sector is disaggregated in many subsectors. Assumption d is indeed very realistic in an aggregate model, although it is no longer so when services and possibly housing are disaggregated.

In spite of the above limitations, it is felt that the analysis of such a simplified model will be useful to understand the basic structure of the relationship between location patterns and space, without overshadowing it with other social and economic details.

Applications of equations (60) yields

$$S^1_{ij} = f^1_{ij}\, e^{-(\nu^1_i - a_{12}\nu^2_j)}$$ the households working in i and living in j

$$S^2_{ij} = f^2_{ij}\, e^{-(\nu^2_i - a_{12}\nu^1_j)}$$ the trips made by households living in i to services in j

$$U^2_i = g^2_i\, e^{-\nu^2_i}$$ the unsatisfied demand for services in i

$$V^1_j = h^1_j$$ the unused housing capacity in j

and substitution of the above results into (64) gives

$$Q^1_j = e^{a_{12}\nu^2_j} \sum_i f^1_{ij} - e^{-\nu^1_i} + h^1_j \qquad (106)$$ the total housing capacity required in j

$$Q^2_j = e^{a_{21}\nu^1_j} \sum_i f^2_{ij}\, e^{-\nu^2_i} \qquad (107)$$ the total service capacity required in j.

Equations (106) and (107) can be introduced in the algorithms of figs. 3 and 4; if the following quantities are defined

z_1 the minimum feasible size for a housing facility

z_2 the minimum feasible size for a service facility

then the heuristic optimality conditions (97) and (98) become

if $Q_j^1 \geq z_1$ establish a housing stock of size Q_j^1

if $Q_j^2 \geq z_2$ establish a service stock of size Q_j^2 in j.

Equations (106) and (107) are very simple, and express the required capacities in terms of the dual variables v_i^1 and v_i^2. If, however, a more «physical» representation is preferred, then equation (104) may be used, and the result is

$$Q_j^1 = \left(\frac{\Phi_j^2}{G_j^2}\right)^{a_{12}} \psi_j^1 + h_j^1 \tag{108}$$

$$Q_j^2 = \left(\frac{\Phi_j^1}{G_j^1}\right)^{a_{21}} \psi_j^2 \tag{109}$$

where

Φ_j^1 is the accessibility to the housing stock for householders working in j

Φ_j^2 is the accessibility to services for households living in j

ψ_j^1 is the potential of housing demand in j

ψ_j^2 is the potential of service demand in j

G_j^1 is the demand for housing generated in j, that is, the number of householders working in j

G_j^2 is the demand for services generated in j, that is, the number of trips to services made by households living in j.

While the interpretation of equation (108) is straightforward, equation (109) requires some explanation. It may seem strange that the total service capacity in j, Q_j^2, is proportional to a power of the accessibility to the housing stock from j, Φ_j^1. But it must be recalled that services need workers, and thus generate housing demand. Equation (109) states a simple balancing principle, by means of which service location is determined both by nearness to demand (by means of demand potential Ψ_j^2), and by nearness to housing facilities required for workers in the service sector (by means of the accessibility Φ_j^1).

4.2. *The health care system*

Let a health care system be given which satisfies the following assumptions.

a. The system consists of N types of facilities, each type numbered from 1 to N. A facility belongs to level k if it is of type k.
b. Patients go to a facility of level $k \neq 1$ either from their residence or from a facility of level $k-1$. Patients go to a facility of level 1 only from their residence.
c. The service capacity is fully used in all levels, that is, the health care demand is congestion insensitive.
d. The minimum feasible size of the facilities of each level is given.

The ordering of health care facilities into levels is usually associated with different degrees and specializations of treatments. For instance, the first level might include general-purpose day-care facilities, usually fairly scattered and accessible; the second level might include urban hospitals, where more specialized and infrequent treatments are available, usually localized in few places; the third level might include regional hospitals, where very specialized treatments are available, usually very localized. The number of levels may vary with different health care organizations in different countries. Assumption b should be relaxed if further disaggregations of specialities within the same level are introduced. In this case a tree structure, rather than a simple ordering, might be more appropriate. However, here only the aggregate case will be considered, in order to keep the example as simple as possible.

The following definitions will be needed

$a_{k,k+1}$ is the fraction of patients in the facilities of level k which require a treatment in a facility of level $k+1$. It will be assumed that $0 < a_{k,k+1} < 1$ for all $k = 1,\dots,N-1$, and $a_{N,N+1} = 0$

Y_i^k is the demand for facilities of level k from the residences in i

z_k is the minimum feasible size for a facility of level k.

Application of equations (60) and (61) yields

$$S_{ij}^k = f_{ij}^k e^{-(\nu_i^k - a_{k,k+1}\nu_j^{k+1})}, \quad \text{for } k \neq N$$

$$S_{ij}^N = f_{ij}\, e^{-\nu_i^N}$$

$$U_i^k = g_i^k\, e^{-\nu_i^k}$$

and substitution of the above results into (64) gives for the required capacities

$$Q_j^k = e^{a_{k,k+1}\nu_j^{k+1}} \sum_i f_{ij}^k e^{-\nu_i^k}, \qquad \text{for } k \neq N \tag{110}$$

$$Q_j^N = \sum_i f_{ij}^N e^{-\nu_i^N}. \tag{111}$$

The heuristic optimality conditions (97) and (98) become

if $Q_j^k \geq z_k$ establish a facility of level k in location j.

A reinterpretation of (110) and (111) by means of equation (104) is also possible. The required service capacities, expressed in terms of accessibilities, potentials, and generated demands, are

$$Q_j^k = \left(\frac{\Phi_j^{k+1}}{G_j^{k+1}}\right)^{a_{k,k+1}} \psi_j^k, \qquad \text{for} \qquad k \neq N \tag{112}$$

$$Q_j^N = \psi_j^N. \tag{113}$$

Equation (112) states that the size of the facility of level k required in location j depends both on the demand potential for level k (a term depending on the residences and the levels below k) and on the accessibility to the facilities of level k + 1. Equation (113) states that the size of the facility of the highest level, N, required in location j depends only on the demand potential for level N, that is, from the residences and the levels below N.

A special case is worth being mentioned, which is relevant for the health care example. Although it has always been assumed that customers behave according to a spatial interaction model, it may be possible that the transport of patients between some levels assumes an emergency character. In this case, the choice of the destination is no longer left to the customer; moreover, it is reasonable to assume that the accessibility sensitiveness disappears, since all emergency cases must be served. One possible approach to introduce emergency trips in the optimization model is as follows. If r is the level to which emergency trips are made, the corresponding terms

$$-\sum_{ij} S_{ij}^r \left(\log \frac{S_{ij}^r}{f_{ij}^r} - 1\right) - \sum_i U_i^r \left(\log \frac{U_i^r}{g_i^r} - 1\right) - \sum_j V_j^r \left(\log \frac{V_j^r}{h_j^r} - 1\right)$$

in the primal objective function (30) are replaced by the single term

$$- \sum_{ij} S^r_{ij} t_{ij}$$

where t_{ij} is the travel time between locations i and j. Moreover, the inequality

$$S^r_{ij} \geq 0$$

is added to the list of constraints. It may be easily shown that the terms $U^r_i(\nu)$ and $Q^r_j(\nu)$ disappear from the dual objective function (65), and the dual problem is no longer unconstrained, since the constraints

$$t_{ij} \geq \sum_k a_{rk} \nu^k_j - \nu^r_i \tag{114}$$

must be met. It may also be shown that the choice of the destination in this case reduces to the nearest-facility rule. Therefore, the required capacity for a facility of level r in location j is no longer given by (64), but by

$$Q^r_j = \sum_i (Q^{r-1}_i a_{rk} + Y^r_i) \, \delta_{ij} \tag{115}$$

where

$$\delta_{ij} = \begin{cases} 1 \,, & \text{if j is the nearest location of a facility of level r} \\ 0 \,, & \text{otherwise.} \end{cases}$$

Another approach to introduce emergency trips is as follows. A very steep function f(t) of travel time may be defined, like the one shown in the graph of fig. 5, and the impedance factor for level r may be given the value:

$$f^r_{ij} = f(t_{ij}).$$

This approach has the advantage of requiring no changes to the original formulation of the problem. However, computational problems may arise from the small numbers introduced by the function f(t) defined above.

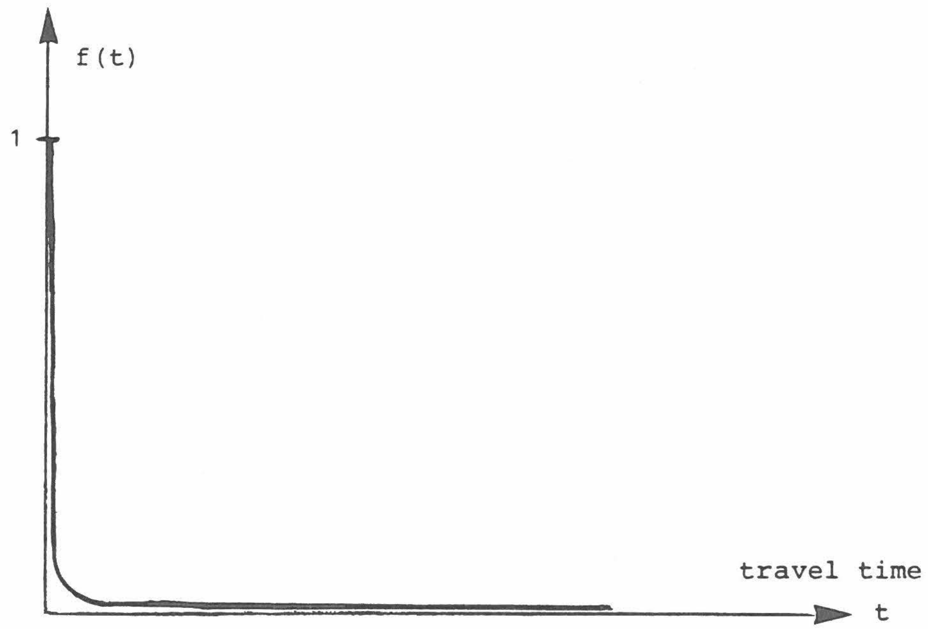


Figure 5 A very steep decay function of travel time

4.3. *The retail system*

Let the multilevel assumption of the health care example be relaxed, so that trips between each pair of activities are possible. The resulting model is an appropriate one for systems where customers make trips with multiple destinations. A typical example is given by a retail system, where customers go shopping for different goods, not necessarily available in the same location. The behavioral model for such a system assumes the most general form discussed in Section 2.2., if the following definitions are introduced

- Y_i^k is the number of trips originating from households living in i and having a retail activity k as a first destination
- a_{rk} is the fraction of customers served in a retail activity r which looks for a retail activity k as the next destination.

If no new assumptions are introduced, the resulting conditions for the optimal sizes are analogous to the ones obtained for the preceding examples. However, the retail system example poses a new interesting problem on the supply side, which is worth being discussed. So far it has been assumed that each activity gives rise to different facilities, and no common costs are shared among them. But in the retail case it may be sensible to lump a subset of different activities together, so as to reach overall economies of scale which each single activity could not reach. This problem may be paraphrased as the problem of the *optimal location and composition of shopping centres.* A slight generalization of the assumptions on establishing costs is required. Let the total cost for establishing a shopping centre in location j be given by the sum of

$a_j^k + b_j^k Q_j^k$	a linear-plus-fixed charge cost to be paid for each activity k
q_j	an overall fixed charge cost to be paid for establishing a shopping centre in j, independently of its size and of the retail sectors it is composed of.

By means of the above assumptions the modified dual objective function (66) can be generalized as follows

$$D(\nu,x,y) = \sum_{jk} [U_i^k(\nu) + \nu_i^k Y_i^k] + \sum_j \{\sum_k x_j^k [Q_j^k(\nu) - \lambda a_j^k] - \lambda q_j y_j\} \qquad (116)$$

where the new variables y_j are defined as

$$y_j = \begin{cases} 1, & \text{if a shopping centre is established in location j} \\ 0, & \text{otherwise.} \end{cases}$$

The resulting saddle-point problem, analogous to (67)-(69), is

$$\max_{x,y} \min_{\nu} D(\nu,x,y) \qquad (117)$$

$$\text{s.t.} \quad 0 \leq x_j^k \leq y_j \qquad (118)$$

$$x_j^k \in \{0,1\} . \qquad (119)$$

The constraints (118) have the following meaning. When at least one activity is established in j, the way the function (116) has been built forces y_i to assume the value 1, that is, a shopping centre is open in j. When no activity is established in j, the $y_i = 0$ and no shopping centre is open in j. A comparison of (117)-(119) with (67)-(69) shows that the constraints (68), requiring no more than one activity in each location, have been dropped. These constraints have already been relaxed in Section 3.3., and for the shopping centre problem they are clearly meaningless, since an optimal combination of *many* different activities in the same location is looked for.

Problem (117)-(119) may be called a «nested» fixed charge problem, since fixed costs have to be paid at two levels, the activity level and the location level. Arguing as for (97)-(98), the following heuristic optimality conditions are obtained

if the subset A_j *of activities for which*

$$Q_j^k \geq \lambda a_j^k$$

is nonempty, and

$$\sum_{k \in A_j} (Q_j^k - \lambda a_j^k) - \lambda q_j \geq 0 \qquad (120)$$

a shopping centre composed of the activities $k \in A_j$ *is established in* j;

if A_j *is empty, no shopping centre is established in* j.

The above conditions may be looked at as a special form of ROT2, where the expression «as many activities as possible» of step 2 has been given a very precise meaning, which is: all activities in the subset A_j, provided it is nonempty and (120) is satisfied. Otherwise stated, activities are required to cover not only their own costs, but also the overall fixed cost for the location they share in common.

5. Concluding comments and issues for further research

As it has been stated many times in this paper, practical tools for urban planning should not place a disproportionate effort in looking for exact solutions to optimization problems. The ill-defined nature of many data and assumptions make it hard to give a realistic meaning to such an exactness. In the field of location problems, it is felt that a better understanding of customer behavior is by far more important than superimposing an «optimal» solution on poor behavioral assumptions. On the contrary, the main effort in the existing literature on location models has been placed on developing hundreds of algorithms to solve problems based on empirically untenable behavioral assumptions. The generalized multiactivity spatial interaction model proposed in Section 2. is not necessarily the best possible one, but it is felt that the model is able to introduce some commonly neglected features of customer behavior, like sensitiveness to accessibility and congestion, in a realistic way. When such features are embedded in an optimization framework, like the one developed in Section 3., the resulting mathematical programs are usually hard to solve. However, there are good reasons to believe that heuristic solutions to a problem based on sound behavioral assumptions are possibly better than exact solutions to a problem based on an over-simplified model of customer behavior. This spirit has guided the development of the methods suggested in Section 3.3. and further specialized in the examples of Section 4. It has also been shown that all the steps of these methods have intuitive interpretations. This may sometimes be dangerous, since common sense and mathematical optimality do not necessarily agree all the time. However, this danger is more than compensated for by the deeper insight which is gained in the structure of the location problem.

Another major goal of this paper has been to provide flexible methods. The outputs of the algorithms proposed in Section 3.3. should be used in a qualitative way, rather than in a quantitative one. The resulting rules for ranking the activities and the locations are much more important than the specific facility sizes resulting from a given set of input data. Sensitivity analysis has also been suggested as a standard approach, and possibly as the only sensible one to solve goal assessment problems.

The above remarks should not be interpreted as an underestimation of the importance of analyzing exact mathematical programming formulations. It has benn shown in Section 3.2. how much insight can be gained simply by looking at the optimality conditions and at some duality relationships. Indeed, the exact optimality conditions and at some duality relationships. Indeed, the exact optimality conditions and the duality results are the roots of the proposed heuristics. Further exploration of the formal properties of the exact formulations is therefore an issue for future research, as long as it will provide better economic interpretations and implementable algorithms.

Many other issues for future research can be listed. Some of them are obviously important, like the development of dynamic versions and the introduction of more complex cost functions and constraints. These themes will be developed in forthcoming papers in this series. Two of them deserve special attention for short term applied developments.

The first one is calibration. The whole framework developed in Section 2. may become useless without an efficient technique for calibrating the many parameters involved in it. The main difficulty lies in the introduction of hardly observable quantities, like the potential demand or the unused capacity. Such calibrating problems have already been solved in the simple case of a single activity (Walsh, Gibberd, 1980), and generalizations to the multiactivity case are under study.

The second one is the introduction of the transport network. If the proposed model is applied to the whole urban system, as suggested in Sections 2.3. and 4.1., then also the traffic conditions are significantly affected, and this must be accounted for in benefit and cost evaluation. Descriptive models combining multiactivity spatial interaction systems and traffic assignment have been already developed by some authors, among them Evans (1976). A normative approach to the same problem, but with the transport network assumed as given, has been proposed by Boyce and LeBlanc (1979). The natural next step is therefore introducing the transport network in the list of decision variables. The complexity of the resulting optimization problem may be discouraging, if an exact algorithm is looked for. However, there are good reasons to believe that easily interpretable heuristics can be developed, along the lines suggested in this paper for the location problem. The use of flexible and qualitative decision rules is even more sensible in this case, since usually a planning authority does not disrupt the existing network. Tools for managing the existing network, and possibly indicating the required changes, are therefore needed, and easily interpretable benefit cost indicators and ranking rules seem to be well suited for this purpose.

References

Boyce D.E., LeBlanc L.J. (1979) Synthesizing Network Equilibrium and Design Models for Location of Urban Activities, Paper presented at the Tenth International Symposium on Mathematical Programming, Montreal, Canada.

Coelho J.D., Williams H.C.W.L. (1978) On the Design of Land Use Plans through Locational Surplus Maximization, *Papers of the Regional Science Association, 40*, 71-85.

Dacey M.P., Norcliffe A. (1976) New Entropy Models in the Social Sciences, 1. Elementary Residential Location Models, *Environment and Planning A, 8*, 299-310.

Erlenkotter D. (1978) A Dual-Based Procedure for Uncapacitated Facility Location, *Operations Research, 14*, 361-388.

Evans S.P. (1976) Some Models for Combining the Trip Distribution and Traffic Assignment Stages in the Transport Planning Process, in Beckmann M., Künzi H.P. (eds.) *Traffic Equilibrium Methods*, Springer Verlag, Berlin, 201-228.

Hansen W.G. (1959) How Accessibility Shapes Land Use, *Journal of the American Institute of Planners, 25*, 73-76.

Jefferson T.R., Scott C.H. (1979) The Analysis of Entropy Models with Equality and Inequality Constraints, *Transportation Research B, 13*, 123-132.

Leonardi G. (1979) Some Mathematical Programming Ideas within a Generalized Spatial Interaction and Activity Framework, in Breheny M.J. (ed.) *London Papers in Regional Science 10: Developments in Urban and Regional Analysis*, Pion, London, 28-47.

Leonardi G. (1980a) On the Formal Equivalence of some Simple Facility Location Models, Working Paper 80-21, IIASA, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Leonardi G. (1980b) A Unifying Framework for Public Facility Location Problems, Working Paper 80-79, IIASA, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Lowry I.S. (1964) A Model of Metropolis, RM-4035-RC, The Rand Corporation, Santa Monica, California.

MacGill S.M., Wilson A.G. (1979) Equivalences and Similarities Between some Alternative Urban and Regional Models, *Sistemi Urbani, 1*, 1, 1-40.

Neuburger H.L.I. (1971) User Benefit in the Evaluation of Transport and Land Use Plans, *Journal of Transport Economics and Policy, 5*, 52-72.

Stewart J.Q. (1948) Concerning «Social Physics», *Scientific American*, 178, 20-23.

Walsh P.K., Gibberd R.W., Development of an Entropy Model for Residential Location with Maximum Zonal Population Constraints, *Environment and Planning A* (forthcoming).

Résumé. Ce papier a deux buts. Le premier est de construire une généralisation du modèle d'interaction spatiale avec double-contrainte, qui tient compte de l'élasticité de la demande par rapport à l'accessibilité et à la congestion, de même qu'aux eventuelles activités multiples.
Dans le paragraphe 2 on montre comment on peut faire ceci, utilisant un principe d'extrème semblable à la maximisation du surplus du consommateur proposé par Neuburger. Ce modèle de comportment général est integré en suite avec une procédure de localisation optimale, qui est présenté avec tous les détails techniques et opératifs (dans le paragraphe 3). Le dernier paragraphe est consacré à la discussion de quelques applications relatives à un système urbain, au système de la santé publique et des services commerciaux.

Riassunto. Lo scopo del presente saggio è duplice. In primo luogo, viene costruita una generalizzazione del modello di interazione spaziale doppiamente vincolato, che tenga conto della elasticità della domanda rispetto all'accessibilità ed alla congestione, nonché di attività multiple. Nel paragrafo 2 è mostrato come ciò possa essere fatto mediante un principio di estremo, equivalente alla massimizzazione del surplus dei consumatori, così come è stato proposto da Neuburger. In secondo luogo, tale modello di comportamento generale viene integrato con una procedura di localizzazione ottimale, che viene presentata nei dettagli tecnici ed operativi. Ciò costituisce l'argomento del paragrafo 3. Nell'ultimo paragrafo (paragrafo 4) vengono delineate e discusse alcune tipiche applicazioni, quali quelle relative ad un sistema urbano, ad un sistema di servizi sanitari e ad un sistema di servizi commerciali.

Sistemi Urbani, 3, 1982, 311-330 Guida Editori, Napoli, Italia

Basic principles of interaction in controlling urban development using entropy models

E.A. Dinitz, Y.A. Dubov, Sh.S. Imelbayev, B.L. Shmulyian
Institute for Systems Studies, 29 Ryleyev Street, 119034 Moscow, U.S.S.R.

Received 30 September 1980, in revised form 12 March 1981

Abstract. The paper focuses on modeling interaction within urban areas as means of controlling urban development. Efficient algorithms for transportation models are developed. One of the methods proposed is a linear approximation of relations between inputs and outputs in the entropy model. Another method is based on minimization of the information measure of deviations from the basic solution. Also, linear approximation methods are used for formulating recommendations pertaining to different characteristics of the transportation system.

Key words: transportation models, entropy, interaction, control inputs.

1. Introduction

Rapid growth of urban population and its mobility leads to a lot of problems related to the development of urban transportation. If cities were absolutely symmetric and homogeneous, their transportation network should clearly have the same properties, therefore urban growth would lead to a simple physical increase in the number of standard segments in the transportation network. In reality, the topography of a city site, the historical background of its development and the necessity to concentrate population for productive activities in small and non-uniformly distributed districts, make problems of urban transportation network design, as well as the allocation of housing and facilities, extremely complex. Bad decisions may lead to unexpected overloads of some parts of the transportation system, since any new route or the construction of a new facility will influence the distribution of passengers' trips over the urban area. One way to estimate the quality of such decisions is to model the transport systems.

A simulation model of intra-urban person trips (Shmulyian, Imelbayev, 1978) is developed using a number of hypotheses. The most important assumption consists of a formal identification of the transportation system with a physical system where passengers are treated as undistinguishable «particles» which choose origins, destinations and transport mode (transit or car) in an independent and stochastic way. Among other hypotheses the assumption that passengers choose shortest routes from origins to destinations is pertinent. The model is based on a «thermodynamical» or entropy maximization approach. The main idea is that the probability of the system's maximal entropy state is so close to unity, that other feasible states are practically impossible.

The maximal entropy state of a transportation system is described by a matrix X with elements x_{ij} equal to numbers of passengers travelling from origin i to destination j:

$$x_{ij} = A_i B_j \nu_{ij} \quad , \tag{1}$$

where ν_{ij} are prior probabilities of choosing pair (i, j) derived from the population distribution curve (see Shmulyian, Imelbayev, 1978), and parameters A_i, B_j are computed from

$$A_i = \frac{P_i}{\sum_j B_j \nu_{ij}} \quad , \qquad B_j = \frac{Q_j}{\sum_i A_i \nu_{ij}} \tag{2}$$

to ensure that total trips from i equal P_i, and total trips to j equal Q_j.

Indices i, j range from 1 to n, where n is the number of urban regions, P_i equals the amount of residents in region i, and Q_j is the number of working places in region j. An overall balance condition,

$$\sum_i P_i = \sum_j Q_j = N \tag{3}$$

where N is the total number of working urban residents, is assumed to hold.

Prior probabilities ν_{ij} are determined accounting for time necessary to move from i to j using the shortest path. These probabilities [and also passenger trips (1)] may be changed by construction or reconstruction of transport routes, since this influences travel time. Therefore we may consider changes in the transportation network as a control input directly influencing outputs from the model. Changes in the population distribution P and facilities allocation Q, as may be seen from (1), (2), also influence passenger trips, so their redistribution should also be considered as control inputs.

Notice that the passenger trips matrix X (1) is in general an intermediary output of the model. Direct outputs are: loads on different routes and transport network links; charts of passenger flows; numbers of passengers changing from underground, railways, buses and so on to underground and railway stations; numbers of passengers leaving origins by different transport modes; and other characteristics. All these characteristics are linear scalar functions $\varphi_k(X)$ of passenger flows x_{ij}.

Therefore the problem of controlling transport system development consists of explaining the interdependencies between control inputs P, Q and ν and model outputs $\varphi_k(X)$. Choice of a particular combination of controls that gives the best balanced change of important indices should be undertaken through interaction between decisionmaker and computer. With this in mind, the following three problems are formulated and solved in this paper.

2. Development of the interaction structure

A special feature of the problem is that the input information of the model (1), (2) is formulated in rather specialized terms which are useless for man-machine interaction. On the other hand terms that people find rather natural and customary such as a widely used name for a part of urban area or aggregated notion of transport route may not be put directly into the model. This is why special interpreting programs linked with a data base system and transforming aggregated data into the model's input information and vice versa must be developed. This interaction system consists of direct interaction programs, programs for data transformation, and software of the transport model.

3. Interaction in real time

After the data from the decisionmaker has been included into the interaction system and the regime of the model's functioning has been determined, sequential computation begins. The times taken to obtain solutions completely determine the lags between formulating or correcting control inputs and getting the corresponding results. These lags depend on the model's dimension and also on which computer programs have been used in this particular computational regime. For instance, the transport model for the city of Moscow includes almost 750 regions, so complete computation takes several hours of processing time. The most bulky programs are the computation of shortest paths trees for all regions, the balancing programs [solution of the system (2)] and programs for allocating passenger flows on the transport network to compute the indices φ_k. So to simulate the real time interaction it is essential to develop an approximate model considering all control inputs as just corrections of a basic solution which has been computed beforehand. However, one faces here specific problems of estimating the accuracy of the results.

4. Analysis of the approximate model

Two ways of computing corrections to the basic solution are considered. The simplest one consists of constructing a linear approximation of the original model. Solution of corresponding linear problems significantly reduces the load on programs such as shortest path computations and balancing programs. Notice that even the linear approximation is of high dimension, so it is necessary to use approximate algorithms, which creates additional errors. Besides, new passenger flows have to be added to base flows through the network, which leads to the use of external data storage devices to compute the corresponding shortest path trees.

The second way is related minimization of the information measure of the difference between a base distribution of passenger flows on the network and those feasible distributions resulting from changes in control inputs (P, Q, v). It is evident that this approach leads to solution of more complex nonlinear problems. But changes of flows are computed directly for network links, therefore the problem may be solved rather quickly using only internal storage devices.

4.1. *Interaction structure*

As was noticed before, the model may be controlled by changing the population distribution, the facilities allocation and the transport network. On the first stage of interaction the control regime is determined by choosing any one or a combination of these control inputs. Such a combination of «population distribution», «facilities allocation» and «network changes» is called a «control regime».

After determining the control regime the stage of designing «control variant structure» is to begin. Here any control input included into the «control regime» has to be decoded.

Direct formation of the control variant structure begins from the output of areas and routes list corresponding to the control regime already determined. This consists of choosing from these lists names of the areas in which population distribution and facilities allocation are to be changed and names of the routes which are to be added to the existing transport network.

The «variant» itself is formed at the same time as its structure, and consists of numerical parameters describing the given structure. For example, for areas included into the variant structure it is necessary to determine changes of population distribution or facilities allocation (as absolute or relative increments) and average speeds of traffic flows on the routes. If necessary, these numbers may be changed while preserving the variant structure (of locations and routes where changes occur).

Thus the control regime determines the list of programs which are to be used for computations under changed input data, the variant structure determines lists of areas where population distribution and facilities allocation may be changed as well as a new network structure, and the variant determines quantitative characteristics of these changes.

As already noted, after a complete description of control inputs new computations for changed input data are needed. For example, if population distribution and facilities allocation are changed, a new balancing [solving system (2)], is needed, and if a new route is introduced, all shortest path trees must be re-computed. However if a new variant is considered as a correction to the base variant, approximate algorithms determining corrections in a short time may be proposed. The structure, description and foundations for these algorithms are given below.

After computations the output is formulated in an interactive way. The following basic output forms are used (see Bolbot *et al.*, 1978): data on departures (arrivals) from a region (to region), flows between urban planning zones and different areas distinguished by transport modes, loads on routes and links of the transport network, changes for different kinds of public transport, usage of underground and railway stations, and also distributions of passenger trips over time distinguished by planning zones.

Important problems to be solved in order to generate efficient man-machine interaction include: development of an algorithm for quick computation of a changed variant and formulation of recommendations for changing control inputs. Main features of these algorithms should be their unisophistication and efficiency in realization. One of the following sections proposes solutions to these problems.

4.2. *Linear computation algorithms*

As stated before a sequence of computation algorithms is determined by the control regime developed through interaction. The following control inputs are possible:

a) changes in population distribution;

b) changes in facilities allocation;

c) changes in transport network.

Any combination of a), b), c) may be considered as a control input too, therefore there are 7 control regimes each of them described by its own computation algorithms.

a) Two situations are possible here: change of population distribution from P to P' preserves the general city-wide population balance, i.e. (3)

holds, or the balance is broken. In the first case the algorithm (a1) may take the following form. Put

$$A_i + \Delta A_i = \frac{P_i + \Delta P_i}{\sum\limits_j (B_j + \Delta B_j) v_{ij}} = \frac{P_i + \Delta P_i}{\sum\limits_j B_j v_{ij}} \cdot \frac{1}{1 + \dfrac{\sum\limits_j \Delta B_j v_{ij}}{\sum\limits_j B_j v_{ij}}}$$

$$= \left(A_i + \frac{\Delta P_i}{\sum\limits_j B_j v_{ij}} \right) \frac{1}{1 + \dfrac{\sum\limits_j \Delta B_j v_{ij}}{\sum\limits_j B_j v_{ij}}} \; .$$

Assuming that the denominator is close to 1 if changes of population distribution are small we obtain:

$$\Delta A_i = \frac{A_i}{P_i} \Delta P_i \tag{4}$$

or

$$A_i' = A_i + \Delta A_i = A_i \left(1 + \frac{\Delta P_i}{P_i} \right) . \tag{5}$$

Substituting (5) into (2) yields:

$$B_j' = B_j + \Delta B_j = \frac{Q_j}{\sum\limits_j A_i' v_{ij}} = \frac{Q_j}{\sum\limits_i A_i v_{ij} + \sum\limits_j \dfrac{A_i \Delta P_i}{P_i} v_{ij}}$$

$$= \frac{Q_j}{\sum\limits_i A_i v_{ij}} \cdot \frac{1}{\left(1 + \dfrac{\sum\limits_i A_i \dfrac{\Delta P_i}{P_i} v_{ij}}{\sum\limits_i A_i v_{ij}} \right)} = B_j \frac{1}{1 + \dfrac{B_j}{Q_j} \sum\limits_i A_i \dfrac{\Delta P_i}{P_i} v_{ij}} \tag{6}$$

$$\simeq B_j \left(1 - \frac{B_j}{Q_j} \sum_i A_i \frac{\Delta P_i}{P_i} v_{ij} \right) = B_j \left(1 - \frac{1}{Q_j} \sum_i x_{ij} \frac{\Delta P_i}{P_i} \right) .$$

Changes of flows are determined from

$$\begin{aligned} x'_{ij} &= x_{ij} + \Delta x_{ij} = (A_i + \Delta A_i)(B_j + \Delta B_j)\,\nu_{ij} \\ &\simeq A_i B_j \nu_{ij} + A_i \Delta B_j \nu_{ij} + B_j \Delta A_i \nu_{ij} \\ &= -\frac{x_{ij}}{Q_j}\sum_i \frac{\Delta P_i}{P_i} x_{ij} + x_{ij}\frac{\Delta P_i}{P_i} + x_{ij} \end{aligned} \tag{7}$$

or

$$\Delta x_{ij} = x_{ij}\left(\frac{\Delta P_i}{P_i} - \frac{1}{Q_j}\sum_i \frac{\Delta P_i}{P_i} x_{ij}\right) . \tag{8}$$

The expressions (8) allow us to avoid the repeated balancing operations. Since the transport network does not change recomputation of shortest paths is also avoided.

A second possibility is that balancing constraint (3) may be violated, because of urban population growth. To estimate the changes we have to alter the facilities allocation Q. Since information on Q is absent this change may be done arbitrarily. We do it in such a way as to decrease the quantity of computations (algorithm a2).

Put $P'_i = \gamma_i P_i$ and $B'_j = B_j$. Then

$$A'_i = \frac{P'_i}{\sum_j B'_j \nu_{ij}} = \frac{\gamma_i P_i}{\sum_j B_j \nu_{ij}} = \gamma_i A_i \tag{9}$$

and

$$Q'_j = B'_j \sum_i A'_i \nu_{ij} = B_j \sum_i \gamma_i A_i \nu_{ij} . \tag{10}$$

Notice that ΔP_i values are determined through interaction in the following way:

$$\Delta P_i = \begin{cases} 0 , & i \notin I_1, \ldots, I_F \\ \dfrac{\Delta_S}{|I_S|} , & i \in I_S \end{cases} \tag{11}$$

where I_S is the list of regions included into area S where a total population change Δ_S is allowed for.

b) If only the facilities allocation is changed, expressions similar to (4)-(11) are used, but A_i and B_j change places, for example

$$\Delta B_j = \frac{B_j}{Q_j} \Delta Q_j \qquad (12)$$

$$\Delta A_i = - \frac{1}{P_i} \sum_j \frac{\Delta Q_j}{Q_j} x_{ij} \; . \qquad (13)$$

Passenger flow changes are computed only for shortest path trees beginning from regions with $\Delta Q_j \neq 0$.

c) Consider transport network change as a result of opening a new route. Data on the urban transport network may be used to determine the matrix of shortest travel times between regions. Prior probabilities of choosing a pari (i, j), ν_{ij}, are determined as

$$\nu_{ij} = \frac{f_k}{n_k} , \quad t_{ij} \in \Delta_k \qquad (14)$$

where

$$f_k = \int_{\Delta_k} f(t)\, dt \qquad (15)$$

is the probability of choosing all pairs (i, j) with $t_{ij} \in \Delta_k$, Δ_k are the segments of trips where travel times change, n_k is the number of pairs (i, j) with $t_{ij} \in \Delta_k$, and $f(t)$ is the population distribution curve.

If a new route is introduced, several rows (and columns) in the shortest times matrix (t_{ij}) are changed; these shortest paths trees (in and outgoing) include links of the new route. Therefore changes in flows are defined on a discrete set of transport network variations. A method of approximate computation is given here.

Since the number of regions is large we set $0\left(\frac{1}{n}\right) \simeq \emptyset$. We assume also that flows variations for different rows and columns of the matrix (t_{ij}) are independent. It follows that we may consider travel cost variations as approximated by changes in only the p-th row and column of the matrix (t_{ij}). Therefore a transport network variation generates a new matrix (t'_{ij}) where

$$\begin{aligned} t'_{ij} &= t_{ij} , \quad i, j = 1, \ldots\ldots, n , \quad i, j \neq P \\ t'_{pj} &\neq t_{pj} , \quad j = 1, \ldots\ldots, n \qquad (16) \\ t'_{ip} &\neq t_{ip} , \quad i = 1, \ldots\ldots, n . \end{aligned}$$

Determining ν'_{ij} is then straightforward. For $i,j \neq P$ changes of ν_{ij} are caused only by changes in n_k [see (14)], since neither segment number Δ_k nor the value f_k change.

By definition, n_k is the number of entries in (t_{ij}) such that $t_{ij} \in \Delta_k$. For the variation (16) $(2n-1)$ values of t_{ip} and t_{pj} would be redistributed. We may assume $n_k = n_k$ and

$$\nu'_{ij} = \nu_{ij} , \quad i,j \neq P . \tag{17}$$

For $i = p$ or $j = p$ the change in ν_{ij} is related to a new number of the segment Δ_k corresponding to the given t'_{ip} and t'_{pj}, i.e.

$$\nu'_{ip} \neq \nu_{ip} , \quad \nu'_{pj} \neq \nu_{pj} .$$

Consider now the A'_i and B'_j multipliers. We have

$$A'_i = \frac{P_i}{\sum\limits_{j \neq p} B'_j \nu_{ij} + B'_p \nu'_{ip}} , \quad i \neq p \tag{18a}$$

$$A'_p = \frac{P_p}{\sum\limits_{j \neq p} B'_j \nu'_{pj} + B'_p \nu'_{pp}} \tag{18b}$$

$$B'_j = \frac{Q_j}{\sum\limits_{i \neq p} A'_i \nu_{ij} + A'_p \nu_{pj}} , \quad j \neq p \tag{18c}$$

$$B'_p = \frac{Q_p}{\sum\limits_{i \neq p} A'_i \nu'_{ip} + A'_p \nu_{pp}} . \tag{18d}$$

It evidently follows that we may put

$$\begin{aligned} A'_i &\simeq A_i , \quad i \neq p \\ B'_j &\simeq B_j , \quad j \neq p . \end{aligned} \tag{19}$$

System (18) has a unique solution approximated by (19) for (18a) and (18c), since these expressions differ from (2) for only one of n numbers ν_{ij} in the denominator. However values A'_p and B'_p differ significantly from A_p and B_p. Designating

$$a = \sum_{i \neq p} A_i \nu'_{ip} , \quad b = \sum_{j \neq p} B_j \nu'_{pj} , \quad c = \nu'_{pp} \tag{20}$$

we get the system for determining A'_p and B'_p:

$$A'_p = \frac{P_p}{a + B'_p c} \quad , \quad B'_p = \frac{Q_p}{b + A'_p c} \qquad (21)$$

which may be solved analytically.

Finally,

$$\begin{aligned} x'_{ij} &= x_{ij} \quad , \quad i,j \neq p \\ x'_{ip} &= A_i B'_p \nu'_{ip} \\ x'_{pj} &= A'_p B_j \nu'_{ip} \ . \end{aligned} \qquad (22)$$

Therefore only row p and column p in the matrix x'_{ij} differ from the initial matrix (x_{ij}). This makes it possible to compute shortest paths and passenger flows for a small number of regions for a given transport network variation.

If we have composite controls ab), ac), bc) and abc) the algorithms are used sequentially. Consider, for example, control input ab). Here the following situations are possible:

1) $\sum_i \Delta P_i = \sum_j \Delta Q_j = 0$.

Then to change the population distribution we use algorithm a1, and to change facilities allocations the algorithm b1, employed for changed population distribution, is then applied.

2) $\sum_i \Delta P_i = \sum_j \Delta Q_j \neq 0$.

First a new facilities allocation Q' is determined. Then the algorithm a2 is used to compute new values $\tilde{A}_i, \tilde{B}_j, \tilde{Q}_j$. Finally b1 is utilized to compute A'_i, B'_j. Here ΔQ_j are determined as $\Delta Q_j = Q'_j - \tilde{Q}_j$.

3) $\sum_i \Delta P_i \neq \sum_j \Delta Q_j$.

For example, $\sum_j \Delta Q_j > \sum_i \Delta P_i$. Introduce new variations of facilities allocation $c\Delta Q_j$, $0 < c < 1$, such that $c \sum_j \Delta Q_j = \sum_i \Delta P_i$. We then have case 2). To distribuite excess working places corresponding to $(1-c)\sum_j \Delta Q_j$ we use the algorithm b2).

4.3. *Justification of the algorithms*

Consider the simplifying assumptions supporting the algorithms a) and b). The most essential one is the assumption that ΔB_j (correspondingly ΔA_i) are small.

From (2) the following system for determining ΔA_i, and ΔB_j is easily derived [see also (6)]:

$$A_i + \Delta A_i = \left(A_i + \frac{\Delta P_i}{\sum_j B_j \nu_{ij}}\right) \frac{1}{1 + \frac{\sum_j \Delta B_j \nu_{ij}}{\sum_j B_j \nu_{ij}}}$$

$$\simeq A_i - \frac{A_i^2}{P_i} \sum_j \Delta B_j \nu_{ij} + A_i \frac{\Delta P_i}{P_i} - A_i^2 \frac{\Delta P_i}{P_i^2} \sum_j B_j \nu_{ij}$$

or, neglecting terms of higher order, we obtain

$$\frac{\Delta A_i}{A_i} + \frac{1}{P_i} \sum_j \frac{\Delta B_j}{B_j} x_{ij} = \frac{\Delta P_i}{P_i} . \tag{23}$$

Similarly for ΔB_j:

$$\frac{\Delta B_j}{B_j} + \frac{1}{Q_j} \sum_i \frac{\Delta A_i}{A_i} x_{ij} = 0 . \tag{24}$$

From (24) we obtain

$$\left|\frac{\Delta B_j}{B_j}\right| = \frac{1}{Q_j} \left|\sum_i \frac{\Delta A_i}{A_i} x_{ij}\right| \tag{25}$$

or

$$\left|\frac{\Delta B_j}{B_j}\right| \leq \max_i \left|\frac{\Delta A_i}{A_i}\right| \frac{1}{Q_j} \sum_i x_{ij} = \max_i \left|\frac{\Delta A_i}{A_i}\right| . \tag{26}$$

Expressing $\dfrac{\Delta B_j}{B_j}$ through $\dfrac{\Delta A_i}{A_i}$ from (24) and substituting into (23) we get the system

$$\frac{\Delta A_i}{A_i} = \frac{\Delta P_i}{P_i} + \frac{1}{P_i} \sum_k \left(\sum_j \frac{x_{ij} x_{kj}}{Q_j} \right) \frac{\Delta A_k}{A_k} \tag{27}$$

or

$$\frac{\Delta A_i}{A_i} = \frac{\Delta P_i}{P_i} + \sum_k a_{ki} \frac{\Delta A_k}{A_k} \tag{28}$$

where $a_{ki} = \dfrac{1}{P_i} \sum_j \dfrac{x_{ij} x_{kj}}{Q_j}$.

It is easy to see that $a_{ki} \geq 0$ and

$$\sum_k a_{ki} = \sum_k \frac{1}{P_i} \sum_j \frac{x_{ij} x_{kj}}{Q_j} = \frac{1}{P_i} \sum_j \frac{x_{ij}}{Q_j} \sum_k x_{kj} = 1 \ .$$

Thus we notice that the system (28) is degenerate, therefore for any number i there exists a solution with $\dfrac{\Delta A_i}{A_i} = 0$. Since $a_{ki} \geq 0$, $\sum_k a_{kj} = 1$, at least one number S_i with $a_{s_i i} \geq \dfrac{1}{n}$ exists. Consider a number i*, such that $\left| \dfrac{\Delta A_{i*}}{A_{i*}} \right| = \max_i \left| \dfrac{A_i}{A_i} \right|$. Then from (28) we obtain

$$\left| \frac{\Delta A_{i*}}{A_{i*}} \right| \leq \left| \frac{\Delta P_{i*}}{P_{i*}} \right| + \sum_k a_{ki} \left| \frac{\Delta A_k}{A_k} \right| = \left| \frac{\Delta P_{i*}}{P_{i*}} \right| + \sum_{k \neq s_i} a_{ki*} \left| \frac{\Delta A_k}{A_k} \right|$$

$$\leq \left| \frac{\Delta P_{i*}}{P_{i*}} \right| + \max_k \left| \frac{\Delta A_k}{A_k} \right| \sum_{k \neq s_{i*}} a_{ki*} = \left| \frac{\Delta P_{i*}}{P_{i*}} \right| + (1 - a_{s_{i*} i*}) \left| \frac{\Delta A_{i*}}{A_{i*}} \right| .$$

Substituting the lower bound $\dfrac{1}{n}$ instead of $a_{s_{i*} i*}$ we get

$$\left| \frac{\Delta A_{i*}}{A_{i*}} \right| \leq \frac{n \left| \Delta P_{i*} \right|}{P_{i*}} \ . \tag{29}$$

Substituting (29) into (26) we have the upper bound for $\frac{\Delta B_j}{B_j}$. Note that this bound is not exact, therefore ΔB_j change little if variations ΔP_i are small; so approximate expressions of the algorithms a1, b1 become justified.

4.4. *Most preferable variations of population distribution and facilities allocation*

Notice that the system (23), (24) allows us to determine variations in flows caused by changes in population distribution. The need to determine population distributions and facilities allocation variations with respect to some goal leads to another problem: the determination of those control inputs which give rise to changes of indices describing the transport system in a desired direction.

As was already noticed, indices describing the transport system are linear functions of flows:

$$\varphi(x) = \sum_{ij} c_{ij} x_{ij} \,. \tag{30}$$

Then

$$\begin{aligned} \Delta\varphi(x) &= \sum_{ij} c_{ij} \Delta x_{ij} = \sum_{ij} c_{ij} x_{ij} \left(\frac{\Delta A_i}{A_i} + \frac{\Delta B_j}{B_j} \right) \\ &= \sum_i \frac{\Delta A_i}{A_i} \sum_j c_{ij} x_{ij} + \sum_j \frac{\Delta B_j}{B_j} \sum_i c_{ij} x_{ij} \\ &= \sum_i \alpha_i \frac{\Delta A_i}{A_i} + \sum_j \beta_j \frac{\Delta B_j}{B_j} \end{aligned} \tag{31}$$

where $\alpha_i = \sum_j c_{ij} x_{ij}$, $\beta_j = \sum_i c_{ij} x_{ij}$.

Put

$$\frac{\Delta B_j}{B_j} = -\frac{1}{Q_j} \sum_k \frac{\Delta A_k}{A_k} x_{kj} \,. \tag{32}$$

Substituting (32) into (31) we get

$$\Delta\varphi(x) = \sum_i \alpha_i \frac{\Delta A_i}{A_i} - \sum_j \frac{\beta_j}{Q_j} \sum_i \frac{\Delta A_i}{A_i} x_{ij} \ . \tag{33}$$

Consider the following set of changes in $\frac{\Delta A_i}{A_i}$:

$$\left(\frac{\Delta A_1}{A_1}, \dots, \frac{\Delta A_n}{A_n}\right) \in \Omega = \left\{ \left|\frac{\Delta A}{A} \in R^n \right| \left|\frac{\Delta A_i}{A_i}\right| \leq \gamma,\ i = 1, \dots, n \right\} \tag{34}$$

where γ is a positive number.

If we ontend to minimize the index $\varphi(x)$, it is necessary to solve the linear programming problem with the objective function (33) and constraints (34). It is easy to see that if $\alpha_i - \sum_j \frac{\beta_j}{Q_j} x_{ij} > 0$, $\frac{\Delta A_i}{A_i} = -\gamma$, and if $\alpha_i - \sum_j \frac{\beta_j}{Q_j} x_{ij} < 0$, $\frac{\Delta A_i}{A_i} = \gamma$.

Now we show what change of control inputs corresponds to such a variation. Summing increments of flows over i we obtain

$$\begin{aligned}\sum_i \Delta x_{ij} &= \sum_i x_{ij} \frac{\Delta A_i}{A_i} - \sum_i \frac{x_{ij}}{Q_j} \sum_k x_{kj} \frac{\Delta A_k}{A_k} \\ &= - \sum_k x_{kj} \frac{\Delta A_k}{A_k} \frac{1}{Q_j} \sum_i x_{ij} + \sum_i x_{ij} \frac{\Delta A_i}{A_i} \\ &= \sum_i x_{ij} \frac{\Delta A_i}{A_i} - \sum_k x_{kj} \frac{\Delta A_k}{A_k} = 0 \ .\end{aligned}$$

This represents a case where facilities allocations remain the same.

Summing flows increments over j we determine changes in population distribution:

$$\begin{aligned}\sum_j \Delta x_{ij} &= \sum_j x_{ij} \frac{\Delta A_i}{A_i} - \sum_j \frac{x_{ij}}{Q_j} \sum_k x_{kj} \frac{\Delta A_k}{A_k} \\ &= P_i \frac{\Delta A_i}{A_i} - \sum_k \left(\sum_j \frac{x_{kj} x_{ij}}{Q_j} \right) \frac{\Delta A_k}{A_k} = \Delta P_i' \ .\end{aligned} \tag{35}$$

Since facilities allocation remained the same, the condition of preserving population balance leads to $\sum_i \Delta P'_i = 0$.

Estimate the sum $\Delta P'_i$:

$$\sum_i \Delta P'_i = \sum_i P_i \frac{\Delta A_i}{A_i} - \sum_i \sum_k \left(\sum_j \frac{x_{kj} x_{ij}}{Q_j} \right) \frac{\Delta A_k}{A_k}$$

$$= \sum_i P_i \frac{\Delta A_i}{A_i} - \sum_k \sum_j \frac{x_{kj}}{Q_j} \frac{\Delta A_k}{A_k} \sum_i x_{ij}$$

$$= \sum_i P_i \frac{\Delta A_i}{A_i} - \sum_k \sum_j x_{kj} \frac{\Delta A_k}{A_k}$$

$$= \sum_i P_i \frac{\Delta A_i}{A_i} - \sum_k \frac{\Delta A_k}{A_k} P_k = 0 \,.$$

It follows that for any variations ΔA the population balance holds, and changes in population distribution over different regions are computed by (35).

To determine the signs of $\Delta P'_i$, introduce the index sets:

$$I = \{ i = 1, \ldots, n \mid \frac{\Delta A_i}{A_i} = \gamma \}$$

$$J = \{ i = 1, \ldots, n \mid \frac{\Delta A_i}{A_i} = -\gamma \} \,. \tag{36}$$

1) $i \in I$

$$\Delta P'_i = \gamma P_i - \sum_{k \in I} \left(\sum_j \frac{x_{kj} x_{ij}}{Q_j} \right) \gamma + \sum_{k \in J} \left(\sum_j \frac{x_{kj} x_{ij}}{Q_j} \right) \gamma$$

$$\geq \gamma \left(P_i - \sum_k \sum_j \frac{x_{kj} x_{ij}}{Q_j} \right) = \gamma \left(P_i - \sum_j x_{ij} \right) = 0 \,.$$

2) $i \in J$

$$\Delta P'_i = -\gamma P_i - \sum_{k \in I} \left(\sum_j \frac{x_{kj} x_{ij}}{Q_j} \right) \gamma + \sum_{k \in J} \left(\sum_j \frac{x_{kj} x_{ij}}{Q_j} \right) \gamma$$

$$\leq \gamma \left(-P_i - \sum_k \sum_j \frac{x_{kj} x_{ij}}{Q_j} \right) = 0 \,.$$

It follows that signs of $\Delta P_i'$ coincide with the signs of $\frac{\Delta A_i}{A_i}$ so the direction of changing population distribution is unambiguously determined by the solution of the extremal problem (33), (34).

Now we can show that for any feasible population distribution change, a non-trivial change of $\left(\frac{\Delta A^1}{A_1}, \ldots, \frac{\Delta A_n}{A_n}\right)$ belonging to Ω exists. Values $\frac{\Delta A_i}{A_i}$ are determined from the solution to (35). It is easy to see that this system is degenerate, therefore it always has a nontrivial solution, if the sum of the right sides is equal to zero, which corresponds to feasible changes in population distribution. Now it suffices to put $\gamma = \max_i \left|\frac{\Delta A_i}{A_i}\right|$.

5. Computations based on the minimal information deviation from the base variant

This method may be used for control inputs a) and b). Its advantages over linear algorithms are less computation effort, and a usage of exact algorithms for minimization of the chosen information index of approximation.

The decrease of computation efforts in this case is achieved by substituting solution of a more complex but less time-consuming problem for allocating changed passengers flows on shortest paths from i to j (these operations are simple but many). The former problem is solved by an iteration method with $O(L)$ operations at each iteration (L - number of network links).

Consider the network (oriented and connected graph without loops) consisting of finite sets of vertices V and links E. Let $e = \{e_{ij}\}$ be the incidence matrix. V corresponds to the sets of origins/destinations I with capacities Q_i, $i \in I$, P_i, $i \in I$, which are given and satisfy the constraint

$$\sum_{j \in I} P_i = \sum_{j \in I} Q_i \; . \tag{37}$$

Network flows $X = \{x_{ij} : x_{ij} > 0 \;\forall i,j \text{ such that } e_{ij} > 0\} \geq 0$ satisfy the constraints

$$\begin{aligned} &\sum_{j \in V} x_{kj} = P_k \,, \quad \sum_{j \in V} x_{ik} = Q_k \,, \quad \forall k \in I \\ &\sum_{j \in V} x_{kj} - \sum_{j \in V} x_{ik} = 0 \,, \quad \forall k \in V \not\subset I \,. \end{aligned} \tag{38}$$

Base variant flows representing some initial basic solution are

$$G = \{G_{ij} : G_{ij} > 0 \ \forall i,j \quad \text{such that} \quad e_{ij} > 0\} \geq 0 .$$

An information measure of the deviation between X and G may be estimated as

$$J(G,X) = \frac{N_1 N_2}{N_1 + N_2} \sum_{ij} \left(\frac{x_{ij}}{N_1} - \frac{G_{ij}}{N_2} \right) \ln \frac{N_2 G_{ij}}{N_1 x_{ij}} \tag{39}$$

(see Kullback, 1964), where $N_1 = \sum_{ij} x_{ij}$, $N_2 = \sum_{ij} G_{ij}$.

Therefore to obtain the flows X which minimize the deviation from G but satisfy new population distribution and facilities allocations (P,Q) we have to minimize J(G,X) under constraints (38).

5.1. *Solution algorithm*

Function (39) is strongly concave on the set $X \geq 0$ and constraints (38) are linear equalities. Therefore this problem has a unique solution and may be solved by the method of Lagrange multipliers. The Lagrange function is

$$L(X,\lambda) = J(G,X) + \sum_{k=|V|+1}^{k=|V|+|I|} \lambda_k \left(\sum_{j \in V} x_{kj} - P_k \right)$$

$$+ \sum_{j \in I} \lambda_k \left(Q_k - \sum_{j \in V} x_{ik} \right) + \sum_{k \in V \not\subset I} \lambda_k \left(\sum_{j \in V} x_{kj} - \sum_{j \in V} x_{ik} \right) . \tag{40}$$

Derivatives $\frac{\partial L}{\partial x_{ij}}$, $\forall i,j$ are:

$$\frac{\partial L}{\partial x_{ij}} = \frac{N_1}{N_1 + N_2} \left[-\ln \frac{N_2 G_{ij}}{N_1 x_{ij}} + \left(1 - \frac{N_1 G_{ij}}{N_2 x_{ij}} \right) \right] + \begin{cases} \lambda_i - \lambda_j & \text{for } i \notin I \\ \lambda_{i+|V|} - \lambda_j & \text{for } i \in I \end{cases} .$$

Putting $\frac{\partial L}{\partial x_{ij}}$ equal to zero we get

$$x_{ij} = \frac{N_2}{N_1} G_{ij} \, e^{-1 + \frac{N_1 G_{ij}}{N_2 x_{ij}} - \frac{N_1 + N_2}{N_2} \begin{cases} \lambda_i - \lambda_j & \text{for } i \notin I \\ \lambda_{i+|V|} - \lambda_j & \text{for } i \in I \end{cases}}$$

or introducing new notions

$$\varphi_{ij}(x_{ij}) = \frac{N_2}{N_1} G_{ij} \, e^{-1+\frac{N_1 G_{ij}}{N_2 x_{ij}}} , \quad \forall i,j$$

$$Z_k = e^{-\frac{N_1+N_2}{N_2}\lambda_k} , \quad k = \overline{1, |I| + |V|}$$

we get

$$x_{ij} = [\varphi_{ij}(x_{ij})/Z_j] \begin{cases} Z_i & \text{for } i \notin I \\ Z_{i+|V|} & \text{for } i \in I . \end{cases} \tag{41}$$

Now substitute (41) into (38)

$$Z_{k+|V|} \sum_{j \in V} \varphi_{kj}(x_{kj}) \cdot \frac{1}{Z_j} = P_k , \quad \forall k \in I$$

$$\frac{1}{Z_k} \left(\sum_{i \in I} \varphi_{ik}(x_{ik}) Z_{i+|V|} + \sum_{i \in V \not\subset I} \varphi_{ik}(x_{ik}) Z_i \right) = Q_k , \quad \forall k \in I$$

$$Z_k \sum_{j \in V} \varphi_{kj}(x_{kj}) \frac{1}{Z_j} - \frac{1}{Z_k} \left(\sum_{i \in I} \varphi_{ik}(x_{ik}) Z_{i+|V|} + \sum_{i \in V \not\subset I} \varphi_{ik}(x_{ik}) Z_i \right) = 0 .$$

Assume that x_{ij}^s are constant. Then expressions (42) are the nonlinear system for unknown coefficients Z_k. This system may be solved by the following iterative process

$$Z_{k+|V|}^{t+1} = \frac{P_k}{\sum_{j \in V} \varphi_{kj}(x_{kj}^s) \frac{1}{Z_j^t}} , \quad \forall k \in I$$

$$Z_k^{t+1} = \frac{\sum_{i \in I} \varphi_{ik}(x_{ik}^s) Z_{i+|V|}^t + \sum_{i \in V \not\subset I} \varphi_{ik}(x_{ik}^s) Z_i^t}{Q_k} \quad \forall k \in I \tag{43}$$

$$Z_k^{t+1} = \left(\frac{\sum_{i \in I} \varphi_{ik}(x_{ik}^s) Z_{i+|V|}^t + \sum_{i \in V \not\subset I} \varphi_{ik}(x_{ik}^s) Z_i^t}{\sum_{j \in V} \varphi_{kj}(x_{ij}^s) \frac{1}{Z_j^t}} \right)^{\frac{1}{2}} \quad \forall k \in V \not\subset I$$

This process is similar to the balancing method (Imelbayev, Shmulyian, 1978) and its convergence follows from Movshoritch (1976).

From (41) we determine new values of X^{s+1}, N_1^{s+1} and repeat the iteration process (43). Computation experiments show that convergence is rather good.

Note that if constraints (3) are violated three cases are possible: 1) $\Delta P_i = 0,\ \forall i$, 2) $\Delta Q_i = 0,\ \forall i$, 3) $\sum_i \Delta P_i \neq \sum_j \Delta Q_j > 0$. In first two cases a similar problem without corresponding constraint (38) is solved. Call these problems A and B, with the initial problem being C. In the third case two problems are solved sequentially: if $\sum_j \Delta Q_j > \sum_i \Delta P_i$ - problem C with $\left\{\Delta P_i, \Delta Q_j' = \Delta Q_j \frac{\sum_i \Delta P_i}{\sum_j \Delta Q_j}\right\}$ and problem a with $\{\Delta Q_j'' = \Delta Q_j - \Delta Q_j'\}$ are solved. If $\sum_j \Delta Q_j < \sum_i \Delta P_i$ problem C with $\left\{\Delta P_i' = \Delta P_i \frac{\sum_j \Delta Q_j}{\sum_i \Delta P_i},\ \Delta Q_j\right\}$ and problem B with $\{\Delta P_i'' = \Delta P_i - \Delta P_i'\}$ represent the necessary steps.

6. Conclusions

Main causes restraining application of entropy models in operative planning of urban systems development are the labor consuming preparation of initial data and decoding of results and the time-consuming nature of computations for any variant using the entropy model. But the majority of control inputs and systems reactions are local. Therefore complete computation for every control input is inefficient.

This paper considers basic principles of interaction in controlling urban development. Formal methods of quick computations for control inputs in transport model are developed. The first method is based on a linear approximation of relations between inputs and outputs realized by the entropy model. The second method is related to minimization of the information measure of deviations from the base variant or basic solution, which is assumed known and already computed. In addition linear approximation methods are used for developing recommendations on forming and changing control inputs to improve different indices of the transport system. When such recommendations are communicated from computer to decision-maker they can help to change control inputs in required directions.

References

Bolbot Yu.K., Korotkov Yu.V., Pahomova J.S., Shapir L.M. (1978) Results of Complex Computation of Passengers Flows, *Transport Systems, 4*, 26-30, Institute for Systems Studies, Moscow (in Russian).

Dubov Yu.A., Imelbayev Sh.S., Shmulyian B.L. (1981) Local Variations in Transport System Entropy Model, *Automatica i Telemekhanica, 7* (in Russian).

Imelbayev Sh.S., Shmulyian B.L. (1978) Modelling of Stochastic Communication Systems, Appendix to Russian translation of Wilson A.G. *Entropy in Urban and Regional Modelling*, Nauka, Moscow (in Russian).

Kullback S. (1964) *Information Theory and Statistics*, Wiley and Sons, New York.

Movshoritch S.M. (1976) Sequential Projecting and Coordinate-Wise Minimization for Complex Programming Problems, *Ekonomica i Matematicheskiye Metodi, 12*, 3, 551-557 (in Russian).

Shmulyian B.L., Imelbayev Sh.S. (1978) Models of Urban Passengers Movements, *Transport Systems, 4*, 7-22, Institute for Systems Studies, Moscow (in Russian).

Résumé. Cet essai a pour objet la modélisation de l'interaction dans les zones urbaines en tant que moyen de contrôle du développement urbain. On développe des algorithmes éfficaces pour des modèles de transport. Une des méthodes proposée est l'approximation linéaire des relations entre les «inputs» et les «outputs» du modèle entropique. Une autre méthode est basée sur la minimization de la mesure d'information des déviations de la solution de base. En outre, les méthodes d'approximation linéaire sont utilisées pour formuler des considérations relatives aux différentes charactéristiques des systèmes de transport.

Riassunto. Questo articolo si incentra sulla modellizzazione dell'interazione entro le aree urbane come mezzo di controllo dello sviluppo urbano. Si sviluppano degli algoritmi efficienti per i modelli di trasporto. Uno dei metodi proposti è quello dell'approssimazione lineare delle relazioni tra gli input e gli output del modello entropico. Un altro metodo è basato sulla minimizzazione della misura di informazione delle deviazioni dalla soluzione di base. Inoltre, i metodi di approssimazione lineare vengono usati per formulare delle considerazioni relative alla diverse caratteristiche dei sistemi di trasporto.

Sistemi Urbani, 3, 1982, 331-369 Guida Editori, Napoli, Italia

Disaggregate models of choice in a spatial context

W. van Lierop, P. Nijkamp

Faculteit der Economische Wetenschappen, Vrije Universiteit, Postbus 7161, De Boelelaan 1105, 1007 MC Amsterdam, The Netherlands.

Received 30 September 1980, in revised form 2 July 1981

Abstract. This paper provides a methodological framework for classifying disaggregate spatial choice models. A distinction is made between deterministic and probabilistic models. Especially the class of probabilistic choice models is discussed in greater detail on the basis of a systematic inventory of the assumptions underlying these models. In this respect, also the class of random utility models for spatial choice behavior is dealt with in a coherent way. After a systematic matrix representation of some main findings, the conclusion is drawn that multinomial logic and probit models may be powerful tools in spatial choice analysis. This conclusion is also illustrated by some numerical experiments.

Key words: disaggregate spatial choice models, deterministic and probabilistic models, gravity and entropy models, constant and random utility theory, multinomial logit and probit models.

1. Introduction

The analysis of spatial allocation and choice patterns in urban and regional systems represents a central issue in regional science and geography. In many traditional analyses, *cost-minimizing* or *utility-maximizing* principles have been used to explain and predict spatial behaviour of people; *economic* decision criteria were pivotal elements in analyzing regional and urban mobility patterns.

Since, however, many industrialized countries have reached a full maturity, the locational and mobility choices are guided by less purely economic-oriented criteria. Many qualitative aspects (such as environmental conditions and the quality of the housing stock) are increasingly influencing spatial choice behaviour. This evokes the need for a broader analysis of spatial choice mechanisms.

Furthermore, the allocation and choice patterns demonstrate an increasing heterogeneity and diversity among spatial actors, so that also the question as to the *scale* of analysis (aggregate versus disaggregate) becomes crucial.

This paper is devoted to a critical survey of *modern choice models*, which are being used nowadays, among others, in attempts to approach spatial allocation and choice patterns in a more realistic way, by means of taking into account the abovementioned qualitative aspects and elements of scale in addition to economic decision criteria. After a methodological introduction, several categories of such choice models are reviewed in a spatial context. Having evaluated these classes of choice models, we will pay more specific attention to two disaggregate choice models, viz. the multinomial logit and the multinomial probit model. The features of the latter pair of models will be illustrated by means of a numerical exercise.

2. Methodological remarks

The methodology of spatial choice analysis can be based on several theoretical frameworks. Two main categories can be distinguished, viz. the *traditional* (mainly neoclassical) and the *behavioural* theories.

The *traditional* approach to spatial choice analysis takes for granted the notions of utility and indifference. Usually, the choice criteria in *micro*-economic decision-making are assumed to be the same for all individuals, though the shape of the individual utility functions is not necessarily equal. Consequently, the same set of attributes of a commodity or of an alternative or the same set of commodities will normally not lead to the same utility for all subjects, while also interpersonal utility comparisons are often not possible. It is clear that the application of this traditional theory to spatial *groups* at a more aggregate level (e.g., social classes, income groups) has until now assumed a uniform utility function for all members of that group.

This traditional approach aims at explaining and forecasting spatial interaction patterns on the basis of classical assumptions of rational behaviour and perfect information. There are several limitations inherent in this approach: consumer interactions (e.g., bandwagon effects) are normally neglected, learning effects are left aside, rational decision-making under perfect information is usually an illusion for spatial choice behaviour (such as migration, commuting and shopping), complementarity of individual choices (such as multi-purpose trips) are usually abandoned and often no insight is given into the distribution of the several alternative choices over the groups, i.e. within the population.

The *behavioural* theories are based on motives and attitudes of decision-makers (or groups); see among others Burnett (1973), Clark, Cadwallader (1973), Downs (1970), Golledge, Brown (1967), Gould (1973), Rushton *et al.* (1967), Saarinen (1976) and Simon (1957). In these theories, such notions as «satisficer» principles, bounded rationality, behavioural environment and the distribution of choices over the population play an important role.

The behavioural thories can be subdivided into 2 main classes, viz. the *revealed preference* approach and the *direct preference* approach.

The first class aims at analyzing *realized*, single and complex choices, by means of information on past behaviour (see among others, Pirie, 1976, and Rushton, 1969, 1971). The basic assumption is that human preferences and aims can be inferred *a posteriori* from the *results* of decisions, i.e. from actual behaviour. If one takes for granted consistency of individual choices (transitivity), a similar ranking of alternatives among individuals, and a sufficiently long time period to define indifference, such an *ex post* analysis can also be used to forecast future behaviour.

The revealed preference approach also has several limitations. Mental processes of consumers are neglected, uncertainties and constraints in choices are mainly left aside, and interaction effects are not taken into account, so that the strong parallel between actual behaviour and internal preference structures is illusive. Furthermore, due to lack of reliable information on actual individual behaviour, the revealed preference approach is often macroscopic (aggregate) in nature.

The direct preference approach analyzes choices and choice-processes *ex ante*, by means of (mainly individual or micro-economic) information on preferences and/or perceptions of choice-makers and alternatives, i.e. it concentrates directly on the *process* of decision making. The information, needed to estimate direct preference models, should be based on questionnaires about attributes of the relevant alternatives and choice-makers with clearly measurable values and preference and/or perception dummies. These real and dummy-values define together the «expected utility» of the choice decisions.

Due to its ability to treat individual preference and perception data this direct approach is usually microscopic (disaggregate). It has the potential to take into account mental images and processes, such as the cognitive perception of commodities and external effects. Furthermore, the direct preference approach may take account of different sets of attributes for individuals and a multiplicity of aims. It is compatible with related areas of knowledge (such as psychology). Consequently, this approach leads to a more integrated economic-psychological framework for the explanation of choice behaviour in general and in socio-geographical space in particular.

Clearly, the direct preference approach is in this respect not yet entirely perfect: constraints on spatial choices emerging from the (behavioural) environment are often still hard to integrate, learning processes and future anticipations are hardly touched upon, while also interaction effects are difficult to integrate.

A more unambiguous evaluation of the various behavioural approaches to analyze choices and choice-processes, than by means of the already mentioned judgements, is very hard to give. This is caused by the fact that the aim of the analysis is not entirely the same for each approach individually. Clearly, one may use methodological, theoretical, logical and empirical criteria such as credibility, plausibility, soundness, or empirical verifiability to judge the various approaches (cf. Nooteboom, 1980), but the operationalization of such concepts depends on the specific aim of the analysis and on the reliability and nature of available data. In fig. 1

we have tried to give a simple illustration of the difference between the revealed preference approach and the direct preference approach.

In fig. 1 the position of the revealed preference approach in block II is clear: it analyzes realized choices by means of information on past decisions. The same holds for the position of the direct preference approach in block IV: it analyzes choice processes, based on interviews

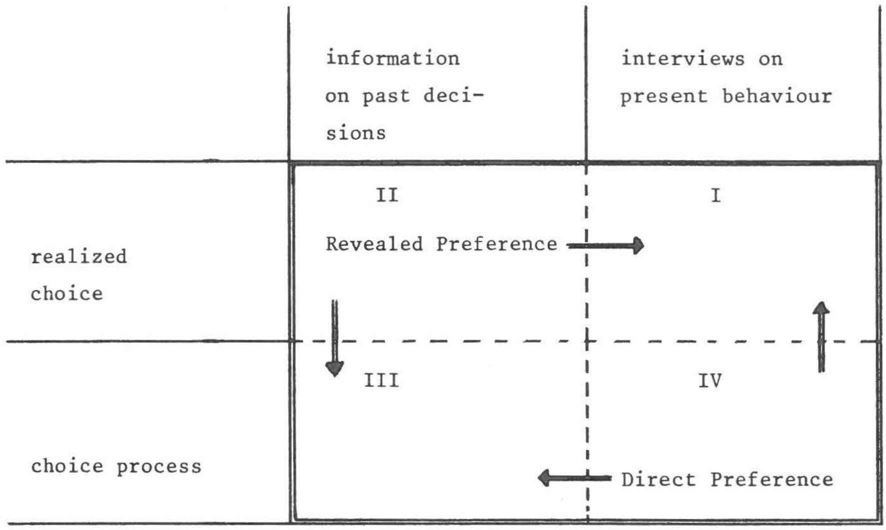


Figure 1 Difference between revealed preference and direct preference approach

about present behaviour. The filling up of the fields I and III is, however, not directly obvious. For instance, those practical analyses of choices and choice-processes which in theory should preferably be done by means of the direct preference approach, are in the estimation phase often forced to use revealed preference data. When these data consist mainly of *recent* ex post data on actual behaviour (for example, in case of repetitive decisions) the influence of the direct preference approach will still be significant. When there exists, however, a bigger time-lag between the depending and explanatory variables in such an estimation model, the revealed preference approach will become more important. The conclusion arises that in blocks I and III of Table 1 interactions take place between the revealed and direct preference *theories* and their more practical model *applications*.

Given the above mentioned features of the various approaches, the behavioural theories and *especially the direct preference approach* appear to us - from a methodological point of view - to be fairly rich in scope (at least compared to the alternative approaches). Further, it has to be stressed that the direct preference approach is rather flexible: it is not at variance with the traditional approach, while some of its elements can also be used in an *ex post* revealed preference approach. However, in order to take better advantage of the direct preference approach, more future research should be done into its mainly practical imperfections. For the improvement of such imperfections recently already attempts have been undertaken to apply so-called «*longitudinal*» analysis of the choice-makers in their decision process (see, among others, Baanders, Slootman, 1980, and Clark, Smith, 1981).

Although the empirical content of behavioural theories of choice, and more specific of direct preference theory compared to revealed preference theory, can only be judged after real world application, it should be noted that there are several techniques (such as multinomial logit and probit models and within their framework the multidimensional scaling techniques) which can help to make this approach potentially useful as an explanatory and forecasting method for measuring choice behaviour.

The rest of the paper will concentrate on a whole set of such techniques applied in a spatial context. The starting point will always be the direct preference theory. It has to be remarked, however, that - mainly due to data problems - many of the described models have regularly been applied as revealed preference models.

3. Direct preference approach and disaggregate models of choice

The direct preference approach assumes that when a certain individual has to choose one alternative from a given set, he will make his choice on the basis of an *ex ante* judgement process about which alternative will presumably give him an *ex post* maximum utility. In psycho-economic modelling this judgement process takes the form of a utility function which may include measured or reported, perceived and preferred values of all kinds of attributes defining the alternative concerned as well as of socio-economic variables characterizing the choice-maker. Such utility functions are used within the framework of a series of choice models that have been developed during the last decade. Our survey study is concerned with a systematic description of such quantitative methods for the analysis of individual, micro or *disaggregate, spatial choice behaviour* in order to get a better understanding of the spatial interactions on a macro or aggregate level.

It has to be noted that, although the philosophy of these methods is entirely concentrated on the choice process of any random individual, sometimes - for reasons of efficiency in the model phases of estimating or forecasting spatial behaviour - the relevant population is subdivided into significantly different groups or classes. Such a classification can be based on features of groups in the population and/or availability of alternatives. In each class the scope of research will then be: what choices will any person - typified as belonging to that class - make? Aggregation over all classes will again give the choice patterns for the entire population.

The description of these disaggregate spatial choice methods is not entirely exhaustive in this paper, but in brief the most recent important developments will be reviewed, with the main *emphasis on model specification*, but occasionally also on model estimation, model evaluation and testing, and aggregation and forecasting, while some attention will also be paid to the coherence of the various model developments.

It has to be noted that the reader is assumed to be familiar with the general foundations, properties and use of disaggregate choice theory and of traditional spatial interaction methods in the form of gravity and entropy - information - models (see, for instance, Stopher, Meyburg, 1975, 1976; Hensher, Stopher, 1979; Manheim, 1979; Nijkamp, 1979, and Van Lierop, Nijkamp, 1980).

Disaggregate models of choice can be divided into two main groups:

A. *Deterministic models*

These are models in which the utility functions are supposed to give an exact description of the alternatives and lagging attributes. Formally, the derivation of the preferred alternative is essentially a mathematically logical conclusion, from which uncertainties - inherent in any choice situation - are eliminated. An aggregation to class or population behaviour presupposes identical utility functions (or identical attributes in the utility functions) concerning the available alternatives for all relevant individuals.

B. *Probabilistic models*

In probabilistic choice theory the choice process of individuals can be described by a straightforward mathematical programming procedure, which defines the probability of a specific choice. This probability decision will partly depend on the observable attributes. In some models also nondirectly measurable elements play an important role.

In other ones, however, the choice probabilities may even be influenced by the mutual relationships between alternatives. In other words, this approach does not, by definition, imply that similar attributes lead to identical choices. The aggregation phase of a probabilistic model requires the definition or approximation of probability density functions for the choice of the available alternatives in the relevant class or in the population.

After a descriptive survey of several models of both groups, we will synthesize in an integrating table form all these models on the bases of their main specification features.

4. Deterministic models

A. *Logit models*

If it is assumed that the relevant utility functions are all *exhaustive* (i.e., all relevant factors are included and exactly measured), we will find among the group of deterministic disaggregated models the well-known logit model, defining in this case the «*choice-ratio*» of any individual n for any alternative i compared to all other alternatives:

$$p_{in} = \frac{\exp u_{in}}{\sum_{i'=1}^{I} \exp u_{i'} n} \; ; \; i,i' = 1,...,I \; ; \; i \neq i' \; ; \; n = 1,...,N \; . \qquad (4.1)$$

with:

p_{in} = «choice»-ratio of individual n for an alternative i compared to all other alternatives (i.e., the ratio in which individual n will «choose» alternative i, out of a set of I alternatives, he has at his availability),

and:

u_{in} = the utility of alternative i for individual n. u_{in} is based on $u(\vec{z}_{in})$, a utility function of a vector $\vec{z}_{in}$ of explanatory, *non-stochastic* variables z_{jin} $(j = 1,...,J)$. J is the number of variables j that define alternative i for individual n. In the deterministic logit model two special assumptions are made concerning $u(\vec{z}_{in})$.

(1) $u(\vec{z}_{in})$ is exhaustive, which means that $\vec{z}_{in}$ exists of exactly measured values for *all* the lagging - both alternative-bounded, and relevant socio-economic - attributes z_{jin}, which are assumed not to be correlated with each other.

(2) Mostly $u(\vec{z}_{in})$ is defined by the following straightforward linear form:

$$u(\vec{z}_{in}) = \sum_{j=1}^{J} z_{jin} \; . \qquad (4.2)$$

Further, the following condition should hold:

$$\sum_{i=1}^{I} p_{in} = 1 \tag{4.3}$$

i.e., the additivity constraints should be satisfied.

A very important hypothesis, which makes the deterministic logit model possible, is the «*independence from irrelevant alternatives axiom*». This property states that the relative *ratio* of «choices» among two particular alternatives is not influenced by introducing other choice options. The advantage of the «independence from irrelevant alternatives» is obviously that the model is easy to handle in terms of estimating and forecasting. A disadvantage is however that when, for instance, a third alternative almost similar to one of the already existing alternatives is taken into consideration as a relevant one, it is very difficult to give a good representation of the various «choice» ratios. This problem may be illustrated by the well-known red-bus blue-bus example of Debreu (1960).

Assume that a certain individual can take the car or a red bus to go to work. The ratios of these traffic modes are 2/3 and 1/3, in other words a ratio of 2:1. Suppose a third alternative, a blue bus, is introduced, which is exactly the same as the first bus except for its colour, but this aspect may be regarded as not of any direct value for the choice between the several modes of travel. Intuitively one would expect as new proportions: 2/3, 1/6, and 1/6. According to the axiom of independence from irrelevant alternatives, however, the ratio of 2:1 between the car and the red bus has to remain the same; consequently, the new choice proportions will become 1/2, 1/4 and 1/4, which gives a wrong idea of reality. Of course this example is rather extreme, but still it illustrates how the calculation of «choice» ratios is influenced in a negative way, when the available alternatives are quite similar. In case the new alternative has features which are completely dissimilar to the initial ones, this disadvantage will not occur.

It should already be mentioned here that it is very difficult - if not impossible - to solve the above described problem completely within the framework of the logit model, because the axiom of independence from irrelevant alternatives is one of the main assumptions of the logit model. Only a step-wise or «nested» logit approach may get round part of the difficulty.

B. *Gravity and entropy models*

Utility based (and in a micro scope developed) gravity and entropy models - which are traditionally more applied to macro situations - can also be regarded as deterministic models of choice (see among others, Nijkamp, 1979). This is no surprise as there exists a direct relationship between the entropy concept from information theory and the logit model (see for more detailed description, van Lierop, Nijkamp, 1979). The basic feature of gravity and entropy models is the assumption that any spatial interaction is the result of three forces: attractiveness at a point of origin, attractiveness at a point of destination and distance friction. On the basis of information on these three forces (and of their related parameters), the spatial flow pattern at the *macro* level can be assessed in a deterministic way. With the aid of that, *micro* «choice» ratios can be derived.

During the last decade these models have gained much popularity in geography and urban and regional science, especially thanks to their relatively simple framework. They have, however, also encountered much criticism for their restrictive underlying assumptions, particularly due to their rather physical concept, with which it is not really possible to *explain* spatial choice behaviour and interaction processes.

5. Probabilistic models

Two main reasons can be formulated supporting the probabilistic approach (see also Andersson, Philipov, 1980). This approach is plausible:

(1) when the attributes are properly observed - i.e., the measurement and/ or perception of all attributes resulted in «realistic», exactly defined, corresponding values, but the decision-making *process* is somewhat stochastic, or the individual does not consistently maximize his utility in terms of neo-classical micro-economic theory (i.e., repeatedly confronted with the same set of exactly described alternatives he will not permanently make the same choices);

(2) when the individual is acting rationally in terms of the neo-classical theory, but when some of the attributes are missing, either just unobserved and/or are observed with a certain bias. In this case the probability approach can be defended on the basis of *observation problems.*

The consequence of these two reasons is the emergence of two directions of theoretical research

1) *the constant utility model,*
2) *the random utility models.*

These approaches regularly take the form of a variation on the basic logit concept. The derivation of the logit model in the constant utility method is, however, different from the one in the more general random utility method. They will therefore both be discussed separately in section 6 and 7-10, respectively.

6. The constant utility model

The basic assumption for this model is the so-called *choice axiom* or IIa problem as defined by Luce (1959). This is the probabilistic version of the hypothesis of independence of irrelevant alternatives in deterministic choice theory. The axiom states that the presence or absence of an alternative is irrelevant to the relative *probabilities* of choice between any two other alternatives, although, of course, the *absolute* values of these probabilities will generally be affected. The same advantages and disadvantages which count for the «independence of irrelevant alternatives rule» are relevant for the choice axiom.

Based on this fundamental assumption Luce (1959) proved that it is possible to define:

$$P_{in} = \frac{M_{in}}{\sum_{i'=1}^{I} M_{i'n}} \; ; \quad i,i' = 1, ..., I \; ; \quad i \neq i' \; ; \quad n = 1, ..., N \; ; \tag{6.1}$$

where:

P_{in} = the choice-probability that individual n wull choose alternative i from a set of I available alternatives. In the constant utility model this choice probability equals the proportion of the population whose choice for alternative i is determined by the same vector of explanatory attributes as the one which is relevant for n. P_{in} is - in contrast with section 4 - written with a capital, because of its stochastic character.

M_{in} = a ratio-number, a cardinal utility figure measured on a scale defined by all available alternatives I, and based on:

(1) a function of the vector of *all* the relevant and perfectly described explanatory variables;

but partly also on:

(2) the analytical way in which the (measured, expected or perceived) utility, resulting from the explanatory variables, is being linked to the choice-probability at the interval $0-1$.

The utility related to point (1) under M_{in} is denoted by u_{in}, and is more or less the same as the comparable figure from the deterministic logit model. Like in this model, the related utility function, $u(\vec{z}_{in})$, is assumed to be exhaustive. In the constant utility model one usually assumes, however, the following - less rigid - linear form for $u(\vec{z}_{in})$:

$$u(\vec{z}_{in}) = \sum_{j=1}^{J} \alpha_j z_{jin} \; ; \tag{6.2}$$

with α_j = a coefficient, a weighing factor.

The relation between the model which describes the *stochastic* process of defining P_{in} and the *deterministic* utility u_{in}, is made by an analytic expression of point (2) under M_{in}; for instance by:

$$M_{in} = e^{u_{in}} . \tag{6.3}$$

Substitution of this into (6.1) provides:

$$P_{in} = \frac{e^{u_{in}}}{\sum_{i'=1}^{I} e^{u_{i'n}}} . \tag{6.4}$$

This description shows that only the *process of choosing* alternative i can be stochastic in the constant utility model, while the utility, u_{in}, itself is fixed and exactly defined.

To illustrate the connection with the logit model framework in a better way, the following statement is made: in fact the above-mentioned means assuming that the probability P_{in} is proportional to an *arbitrary* function G of the vector of the explanatory - alternative related and socio-economic - variables. In formula:

$$P_{in} = \beta_n G(\vec{z}_{in}) , \tag{6.5}$$

with:

β_n = a coefficient of proportionality.

The result of this proportionality assumption is, that the utility function (6.2) does not depend on the actual choice made, as the utility itself is constant! It is predetermined by the exhaustive (fixed) vector of explanatory variables. Taking into account the constraint that:

$$\sum_{i=1}^{I} P_{in} = \sum_{i=1}^{I} \beta_n G(\vec{z}_{in}) = 1 , \tag{6.6}$$

it is easily seen that:

$$\beta_n = \frac{1}{\sum_{i=1}^{I} G(\vec{z}_{in})} , \tag{6.7}$$

or:

$$P_{in} = \frac{G(\vec{z}_{in})}{\sum_{i=1}^{I} G(\vec{z}_{in})} . \tag{6.8}$$

This latter expression is mathematically similar to (6.1) and (4.1). Consequently, formula (6.8) is to be interpreted as a logit model, and the introduction of the assumption of proportionality leads towards the fulfilment of the independence from irrelevant alternatives axiom.

So, the proportionality assumption should be equivalent to the assumption that the choice axiom holds.

7. Random utility approach

Two special features of trying to model individual choice processes by means of a random utility approach are the following:

1) It takes into account that some of the alternative related and/or socio-economic attributes are often missing or unobservable for the researcher. So, these attributes have to be treated as stochastic variables.
2) It assumes that the individual decision-maker chooses an alternative under the condition of bounded rationality. That means, each individual is maximizing his utility rationally, but only in the framework of his personal criteria and his own well defined structure of preferences which is reflected in his personal utility function. Some attributes, which are theoretically assumed to be very important, may not be included into a personal utility function, as they are not preferred by the specific individual or are not perceived as being present. Also the values - personal weighing factors - attached to the attributes can be very different. Consequently, these individual observation elements also lead to a stochastic part in the individual utility function.

A result of these observation problems is that different choices may be made by people having exactly the same set of alternatives, described by exactly the same set of attributes. Consequently, the choice has to be looked upon as a random decision.

Theoretically, the two abovementioned observation problems can then be tackled by assuming that the individual utility or attractiveness, on which the choice for a specific alternative i is based, is a random variable, based on a set of arguments of a *fixed* vector of explanatory variables. This variable is denoted by U_{in}, written with a capital because of its stochastic character. The connected utility function, $U(\vec{z}_{in})$, will be a random function of these explanatory, alternative related and socio-economic, attributes. In formula, $U(\vec{z}_{in})$ can be. represented by:

$$U(\vec{z}_{in}) = v(\vec{z}_{in}) + \eta(\vec{z}_{in}) + \zeta(\vec{z}_{in}) , \qquad (7.1.)$$

in which:

$v(\vec{z}_{in})$ = a deterministic (non-exhaustive) function of the J′ elements of the attribute vector $\vec{z}_{in}$;
in which: $J' \leq J$, the latter representing the maximum number of attributes present. $v(\vec{z}_{in})$ defines the so-called strict (or mean) utility of i for individual n and can be seen as the mathematical expectation of $U(\vec{z}_{in})$. It is the non-exhaustive version of $u(\vec{z}_{in})$ in (6.2). Usually $v(\vec{z}_{in})$ is, equally to $u(\vec{z}_{in})$, assumed to be a linear combination of the perceived or measured values of the fixed set of observed attributes. For all available alternatives, these values may take on any real number. They are assumed to be *normally distributed* and not to be related to each other in any way. As a result $v(\vec{z}_{in})$ can also be assumed to have a normal distribution. It can be written as:

$$v(\vec{z}_{in}) = \sum_{j=1}^{J'} \delta_j z_{jin} : \qquad (7.2)$$

with: δ_j = a coefficient, a weight given to the z_{jin} by the choosing individual n, and
$J' \leq J$ (from the deterministic and constant utility models).

$\eta(\vec{z}_{in})$ = a stochastic function of the following elements:

a. individual «taste variations» over the observed attributes;
b. individual measurement errors, or perception or preference disturbances of the weights, δ_j, given to the attributes, z_{jin}, in (7.2);
c. possible inconsistencies in the individual's choice behaviour, based on $v(\vec{z}_{in})$;
d. the influences of the (restricted) assumption of linearity of the strict utility function.

$\zeta(\vec{z}_{in})$ = a stochastic function of the influence on $U(\vec{z}_{in})$ of missing, omitted or purposely unobserved attributes, (i.e., the impact of the $J-J'$ attributes which might play an important role in the decision process).

As these last two factors are impossible to separate in empirical research, they are always taken together into one stochastic disturbance function $\xi(\vec{z}_{in})$, representing the total individual deviation from the strict utility. Hence, (7.1) should be rewritten into:

$$U(\vec{z}_{in}) = v(\vec{z}_{in}) + \xi(\vec{z}_{in}) \ . \tag{7.3}$$

The at the start of section 3 mentioned maximizing process can in the random utility approach be described as the evaluation process by individual n of the utility U_{in} of all the available alternatives i. In this framework the following general probability statement may be made for any chosen alternative i compared to all other alternatives:

$$P_{in} = \Pr\{U_{in} \geq \max[U_{1n}, \ldots, U_{i-1,n}, U_{i+1,n}, \ldots, U_{In}]\} \ . \tag{7.4}$$

This means in fact:

$$P_{in} = \Pr\{[U_{in} \geq U_{1n}] \wedge \ldots \wedge [U_{in} \geq U_{i-1,n}] \wedge [U_{in} \geq U_{i+1,n}] \wedge \ldots \\ \ldots \wedge [U_{in} \geq U_{In}]\} \ , \tag{7.5}$$

or:

$$P_{in} = \Pr\{U_{in} \geq U_{i'n}\} \ , \quad \forall i' \ . \tag{7.6}$$

In words: the probability that a random alternative i will be chosen by individual n equals the probability that the utility (or attractiveness) of i exceeds or equals the utility of any other alternative i' for n. So it is assumed that the choice maker behaves like a «homo economicus».

By rewritting this last formula with the aid of (7.3), one obtains the *fundamental equation* of the random utility approach:

$$P_{in} = \Pr\{[v(\vec{z}_{in}) + \xi(\vec{z}_{in})] \geq [v(\vec{z}_{i'n}) + \xi(\vec{z}_{i'n})] \ ; \ i, i' = 1, \ldots, I \ ; \\ i \neq i' \ ; \ n = 1, \ldots, N\} \ , \tag{7.7}$$

provided again, of course, that:

$$\sum_{i=1}^{I} P_{in} = 1$$

i.e., the sum of all I choice probabilities should exactly be equal to 1.

The actual calculation of the choice probabilities depends heavily on the *form* that will be chosen for the fundamental equation (7.7). In this respect, the distribution of the error term function, $\xi(\vec{z}_{in})$, plays a crucial role. Several assumptions about this distribution can be made, each leading to different models. In order to present these models in a systematic way, we will follow a classification moving gradually from less to more general models (see also Daganzo, 1979). The following models will be discussed:

(1) models with independent identically distributed error terms;
(2) closed-form models without independent identically distributed error terms;
(3) the multinomial probit model.

These models and their subdivisions will successively be discussed in sections 8-10.

8. Models with independent identically distributed error terms

(A) *Rational Model*

In the rational model (Manheim, 1979), the error terms in formula (7.7) are assumed to be equal to zero. This is useful when the variability of the ξ's is assumed to be small across the given alternatives. A problem is that such a kind of model without error terms is rather unstable: small specification errors may lead to large prediction errors; a change in the attributes of one alternative may lead to a complete shift of the choices.

(B) *Multinomial logit model*

If it is assumed, that the error terms in formula (7.7) are modelled by a set of variates which:

– are mutually independent identically extreme-value distributed (*);

(*) Hensher and Johnson (1981, p. 105), write in this respect: «Independence» indicates that the correlation between the unobserved attributes associated with each and every pair of alternatives in a choice set and across choice sets is zero.
«Identically distributed» says that taste variation exists over the observed attributes (and is allowed for in the random component), yet it is neutral between alternatives, having the same distribution (i.e., equal variance) around the mean (or representative) utility level.

- have zero mean and are uncorrelated with the perceived or directly measured attribute-values and related parameters of the alternatives [i.e., with the parts $v(\vec{z}_{in})$ in (7.7)];
- are consistent with respect to maximization (i.e., in case two disturbances have the same distribution (not necessarily with the same parameters), then also their combined maximum, max $(\xi(\vec{z}_{in}), \xi(\vec{z}_{i'n})$, must have that distribution),

then the so-called multinomial logit (M.N.L.) model results.

These requirements are met by a Gumbel (or Gnedenko or Weibull) distribution, which is a skewed distribution that can be almost normalized by taking logarithms. The Gumbel distribution has the following form:

$$\Pr\{\xi(\vec{z}_{in}) \leq \xi^*\} = e^{-e^{-\xi^*+\gamma}} ; \qquad (8.1)$$

in which: $\xi^* = v(\vec{z}_{i'n}) - v(\vec{z}_{in}) + \xi(\vec{z}_{i'n})$,

and: γ = Euler's constant, $\gamma \simeq 0{,}577$.

When formula (8.1) is assumed to hold it is not difficult to show (for the exact mathematical derivation see McFadden, 1973) that equation (7.7) reduces to the multinomial logit formula:

$$P_{in} = \frac{e^{v(\vec{z}_{in})}}{\sum_{i'=1}^{I} e^{v(z_{i'n})}} . \qquad (8.2)$$

This equation (8.2) is without doubt up till now the most widely used disaggregate demand model to analyse spatial interactions. The reasons for this seem to be that the model was relatively easy to calibrate and that its properties were generally well understood.

One of the most important properties of the M.N.L. model is that, with the assumptions concerning the probability distribution of the error terms, again the same elements are introduced into the model, as defined by the «independence of irrelevant alternatives» hypothesis («choice axiom» or «II-property»). As Holman and Marley proved (see Luce, Suppes, 1965), the choice axiom is theoretically equivalent to the assumptions about the distribution of the error terms in the fundamental equation of the random utility approach. That means that also in the M.N.L. model, the relative probability of choice of two alternatives depends only on their measured attractiveness. (The measurement in this respect might take place directly or indirectly by means of perception or even preference weights, which are transformed into cardinal values).

In cases where the unobserved components of alternatives (the error or rest terms) are correlated, introduction of a new alternative which is highly correlated with another one but is only marginally inferior to it, has hardly any effect on the choice probabilities of all other available alternatives. It is namely highly unlikely that the new alternative will be chosen. When such relations exist, the error terms assumptions of the M.N.L. model cause obvious problems.

Under such circumstances a solution might be to try to capture these interdependences between alternatives by defining adjusted specifications for the functions $v(\vec{z}_{in})$. That is, however, mostly very difficult, as it might mean the introduction of a non-linear function. A few authors have also tried to solve this problem by developing *ad hoc* corrections for the logit model (see among others Domencich, McFadden, 1975); others created new models with interdependent error terms and error terms with different variances. These last mentioned two kinds of models will be reviewed in section 9 after the discussion of a special case of the M.N.L. model in (C), a presentation of a group of different disaggregated sequential choice models in (D), and an overview of a practical *ad hoc* attempt to deal with the choice axiom within the framework of the M.N.L. model in (E) below.

(C) *Binary Logit Model*

A special case of the multinomial logit model - widely used in practice - is the binary logit model. In this choice analysis, it is convenient to express (8.2) as:

$$\ln(P_{1n}/P_{2n}) = v(\vec{z}_{1n}) - v(\vec{z}_{2n}) \, , \qquad (8.3)$$

or, to write:

$$v(\vec{z}_{1n}) - v(\vec{z}_{2n}) = v(\vec{z}_n) \, , \qquad (8.4)$$

so that:

$$\ln\left[\frac{P_{1n}}{1-P_{1n}}\right] = v(\vec{z}_n) \, . \qquad (8.5)$$

The latter expression is traditionally the most common way of representing the binary logit model.

(D) *Elimination By Aspects Method*

Tversky (1972a, 1972b) has developed a disaggregated choice method which makes use of the random utility concept, but is completely different from the methods described sofar. It has mainly been applied in psychology, and does not concern error terms or even assumptions about possible error terms.

The general assumption of this Elimination By Aspects (E.B.A.) Method is that the choice maker selects an alternative in a sequential process based only on the known, identified, explanatory attributes or *aspects* (as they are always called in Tversky's approach). These aspects are scaled in order of importance and are to be interpreted as desirable features. The selection of a particular aspect leads to an elimination of all alternatives which do not contain this desired aspect. The process terminates with the decision based on the last relevant aspect. With that decision the final alternative choice is made.

The analytical description of this procedure yields the Elimination By Aspects model. In the original form of this model, the choice of the aspects, which will be crucial in the successive selection steps, is made at random. The choice probabilities in the model can be defined as an increasing function of the importance of the relevant aspects.

Special features of the E.B.A. model are:

– Aspects which are shared by all alternatives do not affect the final choice probability. This might be a restriction in analyzing interactions, as the relative total values of alternatives cannot exert any influence. The selection process only takes into account the *presence* of aspects. That means that gradual differences of aspects do not have an influence on the choice probabilities. Only completely dissimilar aspects play a role in the E.B.A. model. In analyzing actual spatial interactions this is usually a rigid restriction, as it might exclude *inter alia* relevant socio-economic variables from the decision process. A consequence of this feature is also that the choice between alternatives which are very similar can hardly be explained by a sequential process. Of course this last point is a weak element in most other decision explaining (or forecasting) methods. But in the E.B.A. model there are even in theory no possibilities at all to solve this problem.

– A technical disadvantage of the method is the computational burden which is usually necessary to estimate the outcome of the (assumed) sequential process, and which increases very fast when the number of alternatives and different aspects increase.

Variations of the above mentioned method in which attempts are made at better covering these problems, mainly by a better structuring of the choice process, include:

a. *The Elimination By Tree Method*

This method (see Tversky, Sattath, 1979), assumes a clear tree-structure in its analysis of a choice decision process. All aspects in the process are divided into different branches of a tree, which lead to the available alternatives. There exists, however, no strict hierarchy between the successive choice steps. This means: whenever at random chosen aspect is accepted as a starting point for the evaluation procedure, then one continues the process only in the branches which have been selected by that aspect. The rest of the tree, with all its aspects, is assumed to be not relevant any longer.

b. *The Hierarchical Elimination Method*

This method (see Tversky, Sattath, 1979) is a refinement of the elimination by tree method, in as far it assumes a hierarchy in the choice process. One always starts at the root of the tree and then selects in successive steps the relevant branches via the related aspects.

The technical model description of this method shows a recursive approach in the estimation phase. To be more precisely: in the estimation phase of the model one starts at the level of the realized alternative choices and returns step by step, until the most elementary aspect choice level is reached again. The latter means, until the choice level is reached from which the aspects are the least specific for the ultimate choice, i.e., at the root of the tree.

(E) *Sequential or Multilevel Logit Model*

A practical way of dealing with the choice axiom problem, inherent in the use of the M.N.L. model, is the introduction of sequential or multilevel logit models (see McFadden, 1978). This approach is in line with the above mentioned ideas of assuming a certain hierarchy in choice and decision processes. By splitting the choice problem into several process stages, one will be confronted in the mathematical approach of that process with conditional choice probabilities. The systematics that is introduced in that way into the model to analyze the choice problem can solve part of the difficulties caused by the choice axiom.

This is because, by using such an sequential approach, the number of real choice situations, available alternatives (or sub-alternatives, represented by sets of important explanatory attributes) and the number of parameters in each successive stage, is declining very fast. On the other hand, there will only be a limited loss of efficiency compared to a direct estimation of the model.

A good illustration of a sequential logit model can be given by a spatial distribution problem in which a combined choice has to be made for destination x and mode y.

Assume: $P_{y|xn}$ is the probability that a certain mode will be chosen by an individual n, when his destination is already given, as well as the marginal choice probability P_{xn} for that destination.
So it is supposed that at first x is chosen and only then y.
Furthermore, the assumption should be made that:

$$U(\vec{z}_{xyn}) = A(\vec{z}_{xyn}) + B(\vec{z}_{xn}) \tag{8.6}$$

defines the total utility of the sequentially approached choice problem. The terms at the right hand side of this usually linear function have the following meaning:

$A(\vec{z}_{xyn})$ = a random utility *sub*function based on a vector of directly observed variables or variables which are perceived by person n as important; these variables vary in regard to both destination x and mode choice y;

$B(\vec{z}_{xn})$ = a random utility *sub*function based on a vector of, for person n, important explanatory variables, which vary only in regard to destination x.

Within the framework of the M.N.L. model the result is:

$$P_{y|xn} = e^{U(\vec{z}_{xyn})} / \sum_{y'=1}^{Y_x} e^{U(\vec{z}_{xy'n})}$$

$$= e^{A(\vec{z}_{xyn})} / \sum_{y'=1}^{Y_x} e^{A(\vec{z}_{xy'n})} , \tag{8.7}$$

where:

$y, y' = 1, \ldots, Y_x$; $y \neq y'$, are the possible mode choices for a destination $x = 1, \ldots, X$, for individual $n = 1, \ldots, N$.

Further:

$$P_{xn} = \sum_{y=1}^{Y_x} e^{U(\vec{z}_{xyn})} / \sum_{x'=1}^{X} \sum_{y=1}^{Y_{x'}} e^{U(\vec{z}_{x'yn})}$$

$$= e^{B(\vec{z}_{xn})} \left\{ \sum_{y=1}^{Y_x} e^{A(\vec{z}_{xyn})} \right\} / \sum_{x'=1}^{X} e^{B(\vec{z}_{x'n})} \left\{ \sum_{y=1}^{Y_{x'}} e^{A(\vec{z}_{x'yn})} \right\} , \tag{8.8}$$

in which:

$x, x' = 1, \ldots, X$; $x \neq x'$, are all possible destinations for individual $n = 1, \ldots, N$.

When so-called inclusive values W_{xn} are defined for formula's (8.7) and (8.8) as:

$$W_{xn} = \ln \left\{ \sum_{y=1}^{Y_x} e^{A(\vec{z}_{xyn})} \right\} , \tag{8.9}$$

then they, (8.7) and (8.8), can be written as:

$$P_{y|xn} = e^{A(\vec{z}_{xyn})} / e^{W_{xn}} , \tag{8.10}$$

and

$$P_{xn} = e^{B(\vec{z}_{xn})+W_{xn}} / \sum_{x'=1}^{X} e^{B(\vec{z}_{x'n})+W_{x'n}} . \tag{8.11}$$

These last 2 formula's represent the sequential or multilevel logit model. A method for estimating the joint model, i.e.:

$$P_{xyn} = e^{U(\vec{z}_{xyn})} / \sum_{x'=1}^{X} \sum_{y'=1}^{Y_{x'}} e^{U(\vec{z}_{x'y'n})} , \tag{8.12}$$

is to first estimate the conditional choice model (8.7), then to use that result to define W_{xn} in (8.9), and - after substituting that inclusive value into (8.11) - to estimate finally this last marginal probability model.

A problem with this practically-oriented sequential or multilevel logit model is that, in general, it may be inconsistent with utility maximization (see McFadden, 1978), although it has a good potential to explain real choice behaviour.

9. Closed-form models without independent identically distributed error terms

9.1. *Models with positive correlation between error terms with the same variance*

(A) *Nested Logit Model*

An empirical generalization of the sequential or multilevel logit model is the so-called nested logit model. It is defined by assuming that the inclusive values W_{xn} (from 8.9) have coefficients which are unequal to 1, viz. $(1-\sigma)$; with $\sigma \neq 0$.

The nested logit model then exists of formula (8.10) plus the adjustment of formula (8.11) into

$$P_{xn} = e^{B(\vec{z}_{xn})+(1-\sigma)W_{xn}} / \sum_{x'=1}^{X} e^{B(\vec{z}_{x'n})+(1-\sigma)W_{x'n}} . \qquad (9.1)$$

The special feature of the nested logit model is that it permits *pairwise* correlation between unobserved attributes. For instance, in the example of the physical distribution problem this allows the error term of the step of the destination choice to be more or less correlated with the error term of the phase of the conditional mode choice. So it is possible to take into account special relations between destination and specific mode choices. Yet, at the same time, this «taste variation» is in total still assumed to remain neutral within and between all choice levels of the hierarchically defined process. So all complete alternatives are still assumed to have the same utility distribution for the error terms. In other words: the essential elements of the choice axiom still hold.

The nested logit model has been used quite often in practive (see among others: Ben-Akiva, 1973; McFadden, 1975, 1978; Ameniya, 1976).

(B) *General Extreme Value Model*

McFadden (1978) proved that both the multinomial, the binary, the sequential or multilevel, and the nested logit model are special cases of a family of general extreme value (G.E.V.) choice models. This family is derived from stochastic utility maximization, and allows a *general* pattern of positive correlation among the error terms.

Above all, in the conditional (nested) structure of the models of this family no equal variance of the disturbances is assumed to exist *between* the several levels of the choice process. On the other hand *within* a level the total variances still have to be equal. So within a level the effect of taste variations over the observed attributes in the disturbance term should be neutral in the G.E.V. model.

The *Central Theorem* of the G.E.V. model (see McFadden, 1978, p. 80) is:

suppose that

1. $D(c_1, ..., c_I)$ is a non-negative homogeneous-of-degree-one function of $(c_1, ..., c_I) \geq 0$.
2. $\lim_{c_i \to \infty} D(c_1, ..., c_I) = +\infty$ for $i = 1, ..., I$.
3. For any distinct $(i_1, ..., i_k)$ from $\{1, ..., I\}$, should yield that: $\delta^k D / \delta c_{i_1}, ..., \delta c_{i_k}$ is non-negative if k is odd, and non-positive if k is even.

Then,

$$P_{in} = e^{v(\vec{z}_{in})} D_i(e^{v(\vec{z}_{1n})}, \ldots, e^{v(\vec{z}_{In})}) \, / \, D(e^{v(\vec{z}_{1n})}, \ldots, e^{v(\vec{z}_{In})}) \tag{9.2}$$

defines a probabilistic choice model of alternatives $i = 1, \ldots, I$, which is consistent with utility maximization (see also Daly, Zacharay, 1976).

This formula (9.2) is the basic model of the family of G.E.V. models.

The special case in which:

$$D(c_1, \ldots, C_I) = \sum_{i=1}^{I} c_i \tag{9.3}$$

forms the basis for the M.N.L. model.

From a more general D function satisfying the central theorem of the G.E.V. model:

$$D(\vec{c}) = \sum_{h=1}^{H} a_h \left\{ \sum_{i \in L_h} c_i^{\frac{1}{1-\sigma_h}} \right\}^{1-\sigma_h} \tag{9.4}$$

where:

h (= $1, \ldots, H$) represents the classes in which the alternatives might be grouped,

$L_h \subseteq \{1, \ldots, I\}$,

$\bigcup_{h=1}^{H} L_h = \{1, \ldots, I\}$,

$a_h > 0$,

$0 \leq \sigma_h < 1$,

it is possible to derive the nested logit model with a single class h of the form of formula (8.10) and (9.1), (see McFadden, 1978).

Because of the fact that in (9.4) it is a sufficient condition to satisfy the central theorem when σ lies between zero and one, it is also a sufficient condition for the nested logit model to be consistent with the basic concept of random utility theory, i.e. with stochastic utility maximization. This means that the coefficient $(1-\sigma)$ of the inclusive value in (9.1) has to fall into the unit interval. This coefficient of the inclusive value provides in that case an index of the correlation between the unobserved attributes in the first step of the nested logit model. When $\sigma = 0$ there exists independency, and when σ goes towards 1 the dependency grows. In other words: it can be seen as an indicator of the validity of the choice axiom for the nested logit model.

Functions of the form of (9.4) can also be used to deal with multilevel nested problems, which result into models which can better be described as tree models (see among others: Ben-Akiva, Lerman, 1977; McFadden, 1978; and especially Van Lierop, 1981).

Regarding the empirical possibilities of the G.E.V. method, it has to be mentioned that the qualities of the estimates of the choice probabilities appear to improve compared to describing the actual behaviour with a M.N.L. model. Up till now it is, however, uncertain how efficient this more general model is in cases with more than 3 alternatives.

(C) *Prominence Theory of Choice*

The idea behind this theory, developed by Yu (1979), is that choice alternatives should not be regarded only in terms of their inherent attractiveness, defined by the lagging attributes, but also in terms of their degree of similarity to one another. This means: the relative «prominence» of an alternative becomes a new and relevant choice attribute in any situation. One might regard the figure, which can be introduced as a measure for the prominence of alternatives into the model, describing the choice process, to be a substitute for allowing positive correlation between error terms.

Yu proves empirically, for a case with no more than 3 alternatives, that by taking into account such a new «prominence» attribute, a modified behavioural choice theory is defined within the framework of the logit model and without the strict hypothesis of the choice axioma (IIa assumption).

The model is a variant of McFadden's (1975) elimination by strategy model, which is in turn a generalization of the elimination by aspects method and which introduces into the model - instead of a prominence weight - a similarity factor of alternatives, defined in terms of the degree of overlap of their aspects.

In formula, the prominence theory of choice can be represented by:

$$P_{in} = \frac{e^{v(\bar{z}_{in}) + \log Q_{in}}}{\sum_{i'=1}^{I} e^{v(\bar{z}_{i'n}) + \log Q_{i'n}}}, \tag{9.5}$$

with:

$$Q_{in} = I\left(\sum_{i'=1}^{I} r_{ii'n}\right)^{-1}. \tag{9.6}$$

This term Q_{in} can be regarded as the measure of the average dissimilarity between alternative i and all the other available alternatives, as:

$r_{ii'n}$ = an index of similarity between alternatives i and i' for individual n.

This similarity measure $r_{ii'n}$ can be operationalized among others by defining it as the cosine of the angle (S) between the attribute vector $\vec{z}_{in}$ and $\vec{z}_{i'n}$, i.e.:

$$r_{ii'n} = r(\vec{z}_{in}, \vec{z}_{i'n}) = \cos S\{\vec{z}_{in}, \vec{z}_{i'n}\} = \\ = \frac{\vec{z}_{in}, \vec{z}_{x'n}}{\|\vec{z}_{in}\| \, \|\vec{z}_{i'n}\|} . \tag{9.7}$$

A problem with this most common similarity measure is, however, that it is not necessarily unique. Many other similarity measures can be defined. Maybe this problem can be solved by adding some efficiency criteria.

Yu showed that it is possible to see his prominence theory to be an empirical generalization of both the multinomial logit model and the nested logit model. Especially because of the latter characteristic, this method might also be considered to belong to the family of general extreme value models.

The benefit of the prominence theory of choice-approach, compared to the other so far mentioned models, is that it provides a simple operational variant in cases where attribute dependencies among choice alternatives are relevant, in other words when the choice axiom has to be avoided. However, more study into the similarity measure is necessary in order to resolve the uniqueness problem according to a behaviouristic analysis, and to give the approach also a more firm theoretical foundation.

9.2. *Model with independent exponential distributed error terms*

(D) *Negative Exponential Distribution Model*

In this model (see also Daganzo, 1979, pp. 14-15) the error terms are assumed to be independently, exponentially distributed with a zero mean and different standard deviation $\sigma(\vec{z}_{in})$. So the assumption of «identically» is dropped, compared with (8.1).

In formula:

$$\Pr\{\xi(\vec{z}_{in}) \leq \xi^*\} = \begin{cases} e^{[\xi^* - \sigma(\vec{z}_{in})]/\sigma(\vec{z}_{in})} , & \xi^* \leq \sigma(\vec{z}_{in}) , \\ 1 , & \xi^* > \sigma(\vec{z}_{in}) . \end{cases} \tag{9.8}$$

According to Daganzo (1979), with these assumptions it is possible to obtain the probability of choice after a few algebraic manipulations. Before presenting the choice function, it should be noted that the utility (attractiveness) of any alternative i cannot exceed an upper bound $t(\vec{z}_{in})$, defined by the strict utility and the standard deviation of that alternative for n:

$$t(\vec{z}_{in}) = v(\vec{z}_{in}) + \sigma(\vec{z}_{in}) \ . \tag{9.9}$$

The subscript i indicates the alternative with the ith largest upper bound on the utility as measured and/or perceived by person n, $T(\vec{z}_{in})$.
It has to be assumed that $T(\vec{z}_{0n}) = \infty$ and $T(\vec{z}_{I+1,n}) = -\infty$. Furthermore should hold that $P_{I+1,n} = 0$.

The choice function of the negative exponential distribution (N.E.D.) model can, in the most general case with k classes, then be written as:

$$P_{in} = \sum_{k=1}^{I} \left\{ \frac{\sigma(\vec{z}_{in})^{-1}}{\sum_{i'=1}^{k} \sigma(\vec{z}_{i'n})^{-1}} \cdot \left[\exp\left(- \sum_{i'=1}^{k} \frac{t(\vec{z}_{i'n}) - t(\vec{z}_{kn})}{\sigma(\vec{z}_{i'n})} \right) - \exp\left(- \sum_{i'=1}^{k} \frac{t(z_{i'n}) - t(z_{k+1,n})}{\sigma(\vec{z}_{i'n})} \right) \right] \right\} \ . \tag{9.10}$$

By means of a recursive calibration procedure one can calculate from (9.10) the choice functions for each alternative.
A difficult problem may arise in a model with many alternatives, from which several have approximately the same estimated utility. When in such a case the variance of a distribution term in (9.10) is increased, then at the same time the choice probability of the alternative with the enlarged variance tends to increase.
It might be possible to generalize the N.E.D. model by a model with Weibull error terms. This is a subject of further research.

10. Multinomial probit model

A random utility model in the form of a multinomial probit (M.N.P.) model is characterized by error terms with a joint multivariate normal distribution, with a zero mean and an arbitrary variance-covariance matrix. Thus, in an M.N.P. model the variances of the error terms are allowed to be different and also the error terms themselves are permitted to be mutually correlated. This means that the M.N.P. model can be seen as a generalization of both G.E.V. and N.E.D. models by incorporating the elements of both models.

The starting formula of the M.N.P. model choice function is again the fundamental equation of random utility models as described in chapter 7. By defining a joint multivariate normal density function for the disturbances, $f_{\vec{\xi}(\vec{z}_{\cdot n})}(\vec{\xi}^*)$ with a related distribution function $F_{\vec{\xi}(\vec{z}_{\cdot n})}(\vec{\xi}^*)$, and characterized by:

$$E[\vec{\xi}(\vec{z}_{\cdot n})] = \vec{0} , \qquad (10.1)$$

where:

$$\vec{\xi}(\vec{z}_{\cdot n}) = [\xi(\vec{z}_{1n}), \ldots, \xi(\vec{z}_{in}), \ldots, \xi(\vec{z}_{In})]' , \qquad (10.2)$$

we may formulate the following variance-covariance matrix:

$$E[\vec{\xi}(\vec{z}_{\cdot n}) \cdot \vec{\xi}'(\vec{z}_{\cdot n})] = \Sigma_n . \qquad (10.3)$$

This is equal to:

$$\Sigma_n = \begin{bmatrix} \sigma^2_{1n} & \cdot & \cdot & \cdot & \cdot \\ \sigma_{21n} & \sigma^2_{2n} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{I1n} & \cdot & \cdot & \cdot & \sigma^2_{In} \end{bmatrix} , \qquad (10.4)$$

with:

$$\sigma_{ii'n} = E[\xi(\vec{z}_{in}) \cdot \xi(\vec{z}_{i'n})] , \qquad (10.5)$$

and:

$$\sigma^2_{in} = E[\xi^2(\vec{z}_{in})] . \qquad (10.6)$$

Now, it is possible to estimate the probabilities P_{in}.

For the distribution of the continuous random error term, $\xi(\vec{z}_{in})$, the following characterization is relevant (see, for instance, Mood *et al.*, 1974, p. 138):

$$\Pr[\xi(\vec{z}_{in}) \leq \xi^*] = F_{\xi(\vec{z}_{in})}(\xi^*) \ . \tag{10.7}$$

Introduction of the equivalence of (10.7) viz.: $\int_{-\infty}^{\xi^*} f_{\xi(\vec{z}_{in})}(\xi)\, d\xi$, into the fundamental equation of the random utility approach (7.8) yields the well defined probability function of the M.N.P. model, i.e.:

$$P_{in} = \int_{-\infty}^{\infty} \int_{-\infty}^{v(\vec{z}_{in})-v(\vec{z}_{1n})+\xi(\vec{z}_{in})}, \ldots, \int_{-\infty}^{v(\vec{z}_{in})-v(\vec{z}_{i-1,n})+\xi(\vec{z}_{in})}$$

$$\cdot \int_{-\infty}^{v(\vec{z}_{in})-v(\vec{z}_{i+1,n})+\xi(\vec{z}_{in})}, \ldots, \int_{-\infty}^{v(\vec{z}_{in})-v(\vec{z}_{In})+\xi(\vec{z}_{in})}$$

$$\cdot\ f_{\xi(\vec{z}_{.n})}(\xi(\vec{z}_{1n}), \ldots, \xi(\vec{z}_{in}), \ldots, \xi(\vec{z}_{In}))$$

$$\cdot\ \{ d\xi(\vec{z}_{In}) \ldots d\xi(\vec{z}_{i+1,n})\, d\xi(\vec{z}_{i-1,n}) \ldots d\xi(\vec{z}_{1n})\, d\xi(\vec{z}_{in}) \}$$

$$= \int_{-\infty}^{\infty} F_i(v(\vec{z}_{in}) - v(\vec{z}_{1n}) + \xi(\vec{z}_{in}), \ldots, v(\vec{z}_{in}) - v(\vec{z}_{i-1,n}) + \xi(\vec{z}_{in}),$$

$$\cdot\ \xi(\vec{z}_{in}),\ v(\vec{z}_{in}) - v(\vec{z}_{i+1,n}) + \xi(\vec{z}_{in}), \ldots, v(\vec{z}_{in}) - v(\vec{z}_{In})$$

$$+\ \xi(\vec{z}_{in}))\ d\xi(\vec{z}_{in}) \ . \tag{10.8}$$

With:

$$F_i = \frac{\delta F}{\delta \xi(\vec{z}_{in})} \ .$$

Because of the non-convenient closed form representation of the multiple integrals in (10.8), computation of the choice probabilities of the M.N.P. model was until recently hardly possible. After a first succes by Hausman and Wise (1978), in solving the computational problems of the M.N.P. model, Daganzo (1979) and his associates rediscovered, a single numerical approximation method by Clark (1961), that is surprisingly accurate in estimating M.N.P. models.

This method has the potential to calculate very quickly the choice probabilities for a reasonably large number of alternatives within the framework of a M.N.P. model.

The above mentioned means that the dilemma of the choice axiom (or independence of irrelevant alternatives or IIa hypothesis) can be avoided when one tries to estimate choice probabilities of individual spatial interaction decisions. The possibility to allow a full variance-covariance structure for the random utilities of alternatives in the empirical model-approach of choice processes, which structure permits things like:

- general taste variations between individuals,
- dependence between different alternatives, and
- methods to treat errors in the data in a straighforward manner, leads to the *conclusion* that *the M.N.P. model is a significant realistic disaggregate spatial interaction model.* It has to be remarked, however, that some care should always still be taken with this model in case of many comparable alternatives.

11. Evaluation

In table 2 an integrated survey of the described disaggregate spatial choice models and their most important assumptions is given.

It is clearly seen from sections 3 to 10 and from this table 2 that the disaggregate spatial choice models can be split up into several classes, based on the way in which the relevant models deal with such topics as: dependence between different alternatives, taste variations among individuals and data-errors. The main, although not exclusive, selection criteria in this respect are:

I. *Independence of irrelevant alternatives assumptions*

First one can distinguish the group of models based on the assumption of independence of irrelevant alternatives. The most important representative of this group is the deterministic logit model. This is the least general disaggregate spatial choice model. Implicitly also gravity and entropy models can be regarded as members of this group, at least when these kinds of models can - by means of information theory - be translated into a micro form of the logit model.

Many probabilistic models, the ones which use the choice axiom hypothesis or have independently and identically Gumbel distributed error terms, are influenced by the same features as defined by the independence of irrelevant alternatives assumption.

Models \ features	Independence of irrelevant alternatives	Choice axiom	Error terms independently and identically Gumbel distributed.	Ad hoc solution independence problem	Error terms positive correlated	Error terms independent exponential distributed	Error terms without assumptions
Deterministic Models							
. Logit models	+	0	0	0	0	0	0
. Gravity and Entropy models	-	0	0	0	0	0	0
Probabilistic Models							
- Constant Utility Theory							
. Logit	-	+	0	0	0	0	0
- Random Utility Theory							
+ Models with Independent Identically Distributed Error Terms							
. Rational model	-	-	0	0	0	0	0
. Multinomial logit model	-	-	+	0	0	0	0
. Binary logit model	-	-	+	0	0	0	0
. Elimination by aspects methods	-	-	0	+	0	0	0
1. Elimination by tree model	-	-	0	+	0	0	0
2. 'Hierarchical' Elimination model	-	-	0	+	0	0	0
. Sequential or multilevel logit model	-	-	+	+	0	0	0
+ Closed Form models without Independent Identically Distributed Error Terms							
0 Models with positive correlation between error terms with the same variance							
. Nested logit model	-	-	+	+	+	0	0
. General Extreme Value model	θ	θ	0	0	+	0	0
. Prominence Theory of Choice	0	0	0	+	-	0	0
0 Model with independent exponential distributed error terms							
. Negative exponential distribution model	θ	θ	0	0	0	+	0
+ Multinomial Probit Model	0	0	0	0	θ	θ	+

less general ↔ more general

LEGENDA:
+ = explicit relevant
- = implicit relevant
0 = not relevant
θ = implicit relevant, but more possibilities

Table 2 Integrated survey of disaggregate spatial choice models and their most important assumptions

II. *Choice axiom or IIa hypothesis*

Among the class of models with this assumption one finds the constant utility logit model, and implicitly almost all random utility models, except for some more general ones.

The same model characteristics which are introduced into models by the independence of irrelevant alternatives assumption are found again under the choice axiom or IIa hypothesis.

III. *Assumption of independently and identically Gumbel distributed error terms*

The multinomial, binary, sequential or multilevel and nested logit models fall into this class. This group of models has widely been applied in practice. The assumption of independently and identically Gumbel distributed error terms leads implicitly again to the introduction of the features of the choice axiom or IIa hypothesis.

IV. *Ad hoc solutions independence problem*

It were mainly the shortcomings of the existing models concerning the treatment of such topics as correlation, number of alternatives and individual taste variations, found in the empirical applications, which led to modifications of the well-known disaggregate demand models and to the development of completely new approaches for individual choice analysis. Examples are to be found in the sequential or multilevel logit model, the elimination by aspects method, elimination by tree model and the hierarchical elimination model.

V. *Absence of independent identically distributed error terms in closed form models*

Other approaches try to tackle the problems of independence and taste variations among individuals in a more fundamental way. The general extreme value model may be regarded as the most well-known theory in this respect. It admits the presence of positive correlation among the error terms. The nested logit model and the prominence theory of choice, which can be derived as special cases of the general extreme value model, fall into this class. An alternative approach is offered by the negative exponential distribution model, in which framework independent exponential error terms are defined.

VI. *Presence of full variance-covariance matrix for the rest terms*

The most general theory up till now for real empirical analysis of spatial interaction choices in a disaggregate way is presented by the multinomial probit model. This model, which can intuitively be viewed as the generalization of most of the foregoing methods, allows - as described - the introduction of a full variance-covariance matrix for the rest terms into the individual utility functions *without any restrictive assumptions.* In this way, the M.N.P. model seems not only to be able to meet the required theoretical standards, but also to tackle practice in a very accurate way.

Whether the very high theoretical standards of the M.N.P. model can indeed practically be met and whether they are completely necessary, will depend on the problem at hand. For certain problems the estimation of variances and covariances will be of primary interest to the analyst. Then the model cannot be formulated usefully in an other way than as a M.N.P. model.

In situations in which the set of alternatives is partly unknown or not exhaustively sampled it is, however, even not possible to formulate a M.N.P. model. A model based on the independence of irrelevant alternatives assumption or the choice axiom, like for instance a M.N.L. model, might in such cases still give useful solutions. Sequential and nested structures of spatial interaction choices can up till now in the easiest way be approached with sequential or multilevel and nested logit models.

In case of alternative sampling procedures, like choice-based sampling, the M.N.L. model offers many opportunities. When there are many alternatives available one might have to switch to an elimination by aspects model, etc. Van Lierop and Nijkamp (1980) give a list of methodological, theoretical, logical and practical criteria, which can be helpful when choosing a specific model for a given research project.

Concluding we would like to remark that the M.N.P. model will not necessarily be the only exclusive model of choice for most disaggregate spatial demand studies, although it offers significant theoretical advantages over other disaggregate models of choice in a spatial context.

12. A numerical illustration (*)

As an illustration of the theoretical potential of the M.N.P. model, we *simulated* a spatial choice problem and tested 2 random utility models on it, viz. a M.N.L. and a M.N.P. model. The aim was to show that when one allows for correlation between alternatives belonging to a finite set of spatial choice possibilities, probit model gives in general better results than the logit model. The example is taken from a fictitious homeworkplace commuting situation in the Netherlands. The answers to questions of a home inquiry about individual travel

---

(*) The authors thank A. Rima for her computational assistance.

behaviour have been simulated for a representative group of 250 persons of the relevant population. They had 3 alternatives to go to work: by car, bus or train.

The following model specifications (conform section 7) were used to calculate the choice probabilities:

$$U_{1n} = \gamma_{0\,1n} + \gamma_{1\,1n} z_{1\,1n} + \gamma_{2\,1n} z_{2\,1n} + \gamma_{3\,1n} z_{3\,1n} + \xi(\vec{z}_{1n}) \ , \qquad (12.1)$$

in which:

U_{1n} = the individual utility or attractiveness a random person n attaches to (or expects to get from) alternative 1 of the set of alternatives he has at his availability, here: taking the car to travel from his home to his workplace and back.

$z_{1\,1n}$ = car availability; (the precise meaning of the subscripts is: aspect 1 of alternative 1 for person n).

$z_{2\,1n}$ = valuation of travel time

$z_{3\,1n}$ = costs

$\gamma_{0\,1n}$ = constant term

$\left.\begin{array}{l}\gamma_{1\,1n}\\ \gamma_{2\,1n}\\ \gamma_{3\,1n}\end{array}\right\}$ = parameters

$\xi(\vec{z}_{1n})$ = disturbance factor.

$$U_{2n} = \gamma_{0\,2n} + \gamma_{1\,2n} z_{1\,2n} + \gamma_{2\,2n} z_{2\,2n} + \gamma_{3\,2n} z_{3\,2n} + \xi(\vec{z}_{2n}) \ , \qquad (12.2)$$

in which:

U_{2n} = the utility of individual n in taking the bus to commute between home and work

$z_{1\,2n}$ = valuation travel time

$z_{2\,2n}$ = valuation waiting time

$z_{3\,2n}$ = costs

$\gamma_{0\,2n}$ = constant term

$\left.\begin{array}{l}\gamma_{1\,2n}\\ \gamma_{2\,2n}\\ \gamma_{3\,2n}\end{array}\right\}$ = parameters

$\xi(\vec{z}_{2n})$ = disturbance factor.

$$U_{3n} = \gamma_{13n} z_{13n} + \gamma_{23n} z_{23n} + \gamma_{33n} z_{33n} + \gamma_{43n} z_{43n} + \xi(\vec{z}_{3n}) \;, \tag{12.3}$$

in which:

U_{3n} = the utility of individual n in taking the train to go to work

z_{13n} = valuation travel time

z_{23n} = valuation waiting time

z_{33n} = costs

z_{43n} = valuation comfort

$\left.\begin{matrix}\gamma_{13n}\\ \gamma_{23n}\\ \gamma_{33n}\\ \gamma_{43n}\end{matrix}\right\}$ = parameters

$\xi(\vec{z}_{3n})$ = disturbance term.

An important usual restriction that should be taken into account also with the model in this example is that the sum of the choice probabilities of all the alternatives is exactly equal to 1 and that each of them separately varies between 0 and 1.

It was assumed that the choice utilities for car and bus influence each other and that the choice of the train is independent of the choice of any other alternative. This gave the following variance-covariance matrix:

$$\Sigma_n = \vartheta \begin{bmatrix} T_{1n} & \rho\sqrt{T_{1n}T_{2n}} & 0 \\ \rho\sqrt{T_{1n}T_{2n}} & T_{2n} & 0 \\ 0 & 0 & T_{33n} \end{bmatrix} \;, \tag{12.4}$$

in which:

T_{in} = generalized costs of alternative i ($i = 1, 2, 3$) for individual n ($n = 1, \ldots, n, \ldots, 250$),

ϑ, ρ = parameters which are the same for all alternatives with ranges $-1 \leq \rho \leq 1$ and $\vartheta \geq 0$, which garantees Σ_n to be positive semi-definite.

In the logit model this variance-covariance matrix is of course irrelevant. A direct comparison between multinomial probit and multinomial logit in order to see which one fits better - by means of, for instance, a generalized likelihood ratio test - is not possible, because there is no direct connection between these models: they are not nested, i.e. the one is not a specific case of the other. For a generalized likelihood test it is however a *conditio sine qua non* that the relevant models are nested (*).

To solve this problem a third model should be added as an intermediate one: an independent probit model. This is a special case of the multinomial or covariance probit model and has the following variance-covariance matrix:

$$\Sigma_n = R_{I_n \times I_n} \quad , \tag{12.5}$$

in which:

R = the identity matrix with dimension $I_n \times I_n$ (i.e. defined by the number of alternatives available to individual n).

As a result it is possible to test independent probit and covariance probit against each other.

Hausman and Wise (1978, p. 415) showed that, after normalization of the variances, the independent normal distribution and the extreme value distribution are almost the same. This means that the independent probit model and the logit model are based on approximately the same theoretical aspects (both assume e.g. independence) and also that in practice they will provide about the same results. As a consequence, testing of multinomial probit against independence probit is almost equal to testing multinomial probit against multinomial logit.

The likelihood ratio test itself is based on:

H_0 = independent probit specification with loglikelihood L_0,
H_1 = multinomial probit specification with loglikelihood L_1,
and can be defined as:
reject H_0 if

$$\lambda = -2(L_0 - L_1) > \chi^2_{1-\alpha}(q) \tag{12.6}$$

with: q = number of degrees of freedom
α = probability of an error of type 1.

(*) See Mood *et al.* (1974), p. 419.

The relevant loglikelihood values are calculated with the choice modelling computer program CHOMP, as developed by Daganzo, Schoenfeld (1978). The comparison between the logit and independent probit model is not necessary in this model, because the measurable part of the utility function is identical for both of these specifications, so that also the loglikelihood values will be identical. We found: $L_0 = -305{,}15370$ and $L_1 = -277{,}20543$. This means that the λ from (13.6) becomes: 55,89654. As a result the H_0-hypothesis can be rejected for $\alpha = 0.005$ and $q < 31$ (*). In our example, the number of degrees of freedom is 4 and thus independent probit and also logit will be rejected in favour of covariance or multinomial probit.

From the generated individual choices, one can derive the theoretical *estimations* of the individual probabilities to choose alternative 1, 2 or 3. Putting the mean utility and the variance-covariance into the computer program CONFID, as developed by Sparmann, Daganzo (1979), provides next *predictions* for the multinomial probit case of the probabilities of choosing one specific alternative from the three available options. Also the predictions in case of a logit specification can be calculated by means of that computer program. The prediction results of both are given in Table 3.

Table 3 Individual choice probability *prediction* results of the CONFID computer program

Probabilities	M.N. Probit	Independent Probit	M.N. Logit	Theoretical
alt. 1	0.32311	0.28592	0.38367	0.328
alt. 2	0.35132	0.39079	0.35982	0.356
alt. 3	0.32059	0.31858	0.25651	0.316

In this short example the differences between M.N.L. and M.N.P. are not yet very significant. It seems however justifiable to conclude that on the whole the multinomial probit model gives better results. It is obvious that especially the ratio between alternatives 1 and 2 is very well predicted by the multinomial probit method, due to the existing correlation between the alternatives.

How this kind of probit analysis will work in more complex practical cases, is a new field of research. Results of work recently done are very promising. As a final remark of this illustrative section and a provisional conclusion of the paper we can therefore state that in a case of correlation among alternatives in a spatial choice analysis problem, multinomial probit models are worthwhile to use.

(*) See Daganzo (1979), p. 208.

References

Ameniya T. (1976) Specification and Estimation of a Multinomial Logit Model, Technical Report 211, Institute of Mathematical Studies in the Social Sciences, Stanford University, Stanford, CA.

Andersson Å.E., Philipov D. (1980) Economic Models of Migration, Paper presented at the Regional Development Task Conference on «Theoretical and Practical Aspects of Regional Development Modeling», IIASA, Laxenburg, Austria, March 18-21.

Baanders A., Slootman K.C.P. (1980) Een Panel voor Longitudinaal Verplaatsingsonderzoek, in Bovy P., Le Clerq F., Van Est J., Jansen B. (eds.) *Mobiliteit, Ruimtelijke Interaktie en Vervoerssysteemgebruik*, Colloquim Vervoersplanologisch Speurwerk, Den Haag.

Ben-Akiva M.E. (1973) Structure of Passenger Travel Demand Models, Ph. D. Diss., M.I.T., Department of Civil Engineering, Cambridge, Mass..

Ben-Akiva M.E., Lerman S. (1977) Disaggregate Travel and Mobility Choice Models and Measures of Accessibility, Paper presented at the Third International Conference on Behavioural Travel Modelling, Tanenda, Australia.

Burnett P. (1973) The Dimensions of Alternatives in Spatial Choice Processes, *Geographical Analysis, 5*, 181-204.

Clark C.E. (1961) The Greatest of a Finite Set of Random Variables, *Operations Research, 9*, 145-162.

Clark W.A.V., Cadwallader M.T. (1973) Residential Preferences: An Alternative View of Intra-Urban Space, *Environment and Planning, 5*, 693-703.

Clark W.A.V., Smith T.R. (1981) Expected Utility Theory and Housing Market Search, *Environment and Planning* (forthcoming).

Daganzo C. (1979) *Multinomial Probit, the Theory and its Application to Demand Forecasting*, Academic Press, New York.

Daganzo C., Schoenfeld L. (1978) CHOMP User's Manual, Research Report UCB-ITS-RR-78-7, Institute of Transportation Studies, University of California, Berkeley.

Daly A., Zachary S. (1976) Improved Multi-Modal Split Models, Paper PTRC, Summer Annual Meeting, Warwick.

Debreu G. (1960) Review of Luce R., Individual Choice Behavior, *American Economic Review, 50*, 186-188.

Domencich F.A., McFadden D. (1975) *Urban Travel Demand, a Behavioural Analysis*, North-Holland, Amsterdam.

Downs R.M., Geographic Space Perception: Past Approaches and Future Prospects, *Progress in Geography, 2*, 65-108.

Golledge R.G., Brown L.A. (1967) Search, Learning and the Market Decision Process, *Geografiska Annaler, 49 B*, 116-124.

Gould P.R. (1973) On Mental Maps, Image and Environment, in Downs R.M., Stea D. (eds.) *Cognitive Mapping and Spatial Behavior*, Aldine Press, Chicago.

Hausman J.A., Wise D.A. (1978) A Conditional Probit Model for Quantitative Choice: Discrete Recognizing Interdependence and Heterogeneous Preferences, *Econometrica, 46*, 2, 403-426.

Hensher D.A., Stopher P.R. (1979) (eds.) *Behavioural Travel Modelling*, Croom Helm, London.

Hensher D.A., Johnson L.W. (1981) *Applied Discrete Choice Modelling*, Croom Helm, London.

Lierop van W. (1981) Towards a New Disaggregated Model for the Housing Market, in Lierop van W., Nijkamp P. (eds.) *Locational Developments and Urban Planning*, Sijthoff and Noordhoff, The Netherlands, 181-219.

Lierop van W., Nijkamp P. (1979) A Utility Framework for Interaction Models for Spatial Processes, *Sistemi Urbani, 1*, 1, 41-64.

Lierop van W., Nijkamp P. (1980) Spatial Choice and Interaction Models: Criteria and Aggregation, *Urban Studies, 17*, 299-311.

Luce R.D. (1959) *Individual Choice Behaviour*, Wiley, New York.

Luce R.D., Suppes P. (1965) Preference, Utility and Subjective Probability, in Luce R.D., Bush R.P., Galanter E. (eds.) *Handbook of Mathematical Psychology*, vol. 3, Wiley, New York.
Manheim M.L. (1979) *Fundamentals of Transportation Systems Analysis*, vol. 1: *Basic Concepts*, M.I.T. Press, Cambridge, Mass..
McFadden D. (1973) Conditional Logit Analysis of Qualitative Choice Behavior, in Zarembka P. (ed.) *Frontiers in Econometrics*, Academic Press, New York.
McFadden D. (1975) Economic Applications of Psychological Choice Models, Working Paper 75-19, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley.
McFadden D. (1978) Modelling the Choice of Residential Location, in Karlqvist A., Lundqvist L., Snickars F., Weibull J.W. (eds.) *Spatial Interaction Theory and Planning Models*, North Holland, Amsterdam.
Mood A.M., Graybill F.A., Boes D.L. (1974) *Introduction to the Theory of Statistics*, Third Edition, McGraw-Hill, Kogakusha LTD, Tokyo.
Nijkamp P. (1979) Gravity and Entropy Models: the State of the Art, in Jansen G.R.M., Bovy P.H.L., Van Est J.P.J.M., Le Clerq F., *New Developments in Modelling Travel Demand and Urban Systems*, Saxon House, England, 281-319.
Nooteboom B. (1980) *Retailing: Applied Analysis in the Theory of the Firm*, Gieben, Amsterdam-Uithoorn.
Pirie G.H. (1976) Thoughts on Revealed Preference and Spatial Behaviour, *Environment and Planning A, 8*, 947-955.
Rushton G., Golledge R.G., Clark W.A.V. (1967) Formulation of a Normative Model for the Spatial Allocation of Grocery Expenditures by a Dispersed Population, *Annals, Association American Geographers, 57*, 389-400.
Rushton G. (1969) Analysis of Spatial Behavior by Revealed Space Preference, *Annals, Association American Geographers, 59*, 391-400.
Rushton G. (1971) Behavioural Correlates of Urban Spatial Structure, *Economic Geography, 47*, 49-58.
Saarinen T.F. (1976) *Environmental Planning, Perception and Behavior*, Houghton Mifflin, Boston.
Simon H.A. (1957) *Models of Man*, Wiley, New York.
Sparmann J.M., Daganzo C. (1979) CONFID User's Manual, Research Report UCB-ITS-RR-79-5, Institute of Transportation Studies, University of California, Berkeley.
Stopher P.R., Meyburg A.H. (1975) *Urban Transportation Modelling and Planning*, Lexington Books, Lexington, Mass..
Stopher P.R., Meyburg A.H. (eds.) (1976) *Behavioural Travel Demand Models*, Lexington Books, Lexington, Mass..
Tversky A. (1972a) Choice by Elimination, *Journal of Mathematical Psychology, 9*, 341-367.
Tversky A. (1972b) Elimination by Aspects: a Theory of Choice, *Psychological Review, 79*, 281-299.
Tversky A., Sattath S. (1979) Preference Trees, *Psychological Review, 86*, 542-573.
Yu W. (1979) A Prominence Theory of Choice, Working Paper 10, Regional Science Department, University of Pennsylvania, Philadelphia.

Résumé. Cet essai présente un cadre de référence méthodologique pour classifier des modèles de choix spatial désagregés. On fait une distinction entre modèles déterministes et modèles probabilistes. Ces derniers, en particular, sont étudiés en détail sur la base d'une analyse systématique des suppositions qui les régissent. De ce point de vue on discute aussi, d'une façon cohérente, des modèles d'utilité qui influent sur de comportment de choix spatial. Après avoir construit une representation matricielle systématique des principaux résultats, on arrive à la conclusion que les modèles logit multinomiaux et les modèles probit peuvent être des instruments très efficaces pour l'analyse du choix spatial. Cette conclusion est illustrée par quelques exemples numeriques.

Riassunto. Questo articolo fornisce un quadro di riferimento metodologico per la classificazione dei modelli di scelta spaziale disaggregati. Si distingue tra modelli deterministici e modelli probabilistici. Questi ultimi, in particolare, sono analizzati approfonditamente sulla base di un inventario sistematico delle assunzioni che li governano. Da questo punto di vista, viene discussa anche, coerentemente, la classe dei modelli di utilità casuale per il comportamento di scelta spaziale. Dopo aver costruito una rappresentazione matriciale sistematica dei principali risultati, si perviene alla conclusione che i modelli logit multinomiali ed i modelli probit possono essere degli strumenti efficaci nell'analisi della scelta spaziale. Questa conclusione viene anche illustrata con alcuni esempi numerici.

Sistemi Urbani, 3, 1982, 371-401 Guida Editori, Napoli, Italia

A multilevel economic-demographic model for the Dortmund region

M. Wegener

Institute of Urban and Regional Planning, University of Dortmund, August-Schmidt-Strasse 10, 46 Dortmund 50, F.R.G..

Received 30 September 1980, in revised form 10 March 1981

Abstract. The paper is a status report on a research project conducted at the Institute of Urban and Regional Planning of the University of Dortmund in which the relationships between economic, i.e. sectoral and technological, change, locational choice, mobility, and land use in urban regions are investigated. For this purpose a multilevel dynamic simulation model of regional development was designed to simulate location decisions of industry, residential developers, and households, the resulting migration and commuting patterns, the land use development, and the impacts of public programs and policies in the fields of industrial development, housing, and infrastructure.
The model consists of three spatial levels: (1) a macroanalytic model of the economic and demographic development of 34 labor market regions in the state of Nordrhein-Westfalen, (2) a microanalytic model of intraregional location and migration decisions in 30 zones of the urban region of Dortmund, and (3) a microanalytic model of land use development in one or more districts of Dortmund.
The paper contains a description of the general structure of the model system and its major levels and submodels as well as a summary of the data collection and calibration problems encountered when implementing it for the urban region of Dortmund.

Key words: urban change, migration, housing, multilevel simulation.

0. Introduction

In the development of settlement systems in the Federal Republic of Germany three phases can be distinguished. During the fifties employment and population increased in nearly all parts of the country. In the sixties the mechanization of agriculture combined with increasing job opportunities in the urbanized regions led to a massive migration from the rural regions into a few major agglomerations. In the seventies, however, this large-scale *agglomeration* process has been outbalanced by a small-scale *deglomeration* process caused by increased migration of population and industry from the centers of the agglomerations to their less urbanized peripheries.

In combination with other tendencies characteristic for highly industrialized countries - falling birth rates and loss of employment in the primary and secondary sectors of the economy - these spatial shifts result in a decline of population in all larger agglomerations and a decline of employment in some of them. The latter is characteristic for old, monostructured industrial regions like the *Ruhr* in which employment losses in the secondary sector are not compensated by gains in the tertiary sector.

Is this the end of the century-old urbanization process? Regional Science has offered only vague answers to this question. They range from conjectures on «the complete ruin for the 20th century urban culture» (van den Berg, Klaassen, 1978) to the hardly more attractive vision of a «continuous megalopolis from Merseyside to Milan» (Hall, 1977). However, there is still lack of a *theory* of spatial behavior of industry and households by which *simultaneous* processes of agglomeration and deglomeration occurring in one country or even in one agglomeration can be explained in agreement with empirically founded economic and social theories.

One approach to contribute to such a theory sets out from the obvious relation between urbanization and industrialization observed already in the 19th century when the transition from agricultural to industrial society found its spatial expression in the rapid growth of cities. Can this explanatory model of the *urbanization* process be augmented to capture also the *suburbanization* process, i.e. can the present-day process of spatial change be explained as a consequence of its economic context?

The research project reported in this paper is an attempt to supply some answers to this question. It starts from two general hypotheses about the relationships between economic, i.e. sectoral and technological, change, locational choice and mobility of industry and households, and land use in urban regions:

a. the interregional agglomeration process will continue with the process of innovation and structural change of the economy, as industries with high innovation and growth potential benefit more from agglomeration economies than others and will continue to do so;

b. the intraregional deglomeration process is the reverse side of the interregional agglomeration process as it represents the response of enterprises and households to growing diseconomies and the deteriorating quality of life of the agglomeration cores. However, it is also caused by demographic changes and changing life styles and is facilitated by rising incomes and the ubiquity of transport and other infrastructure.

In the light of these hypotheses in particular three questions are to be investigated:

1. What are the impacts of economic change, i.e. intersectoral shifts and technological developments, on the locational choice of enterprises and households in urban regions and the resulting migration and commuting patterns?

2. What are the impacts of the spatial structure of an urban region on the process of economic change in the region?

3. What are the impacts of programs and policies of local government on the locational choice of private actors?

For this purpose a multilevel dynamic simulation model of regional development was designed to simulate location decisions of industry, residential developers, and households, the resulting migration and commuting patterns, the land use development, and the impacts of public programs and policies in the fields of industrial development, housing, and infrastructure.

It was decided to use the urban region of Dortmund as a study region, including Dortmund (pop. 610,000) and 19 neighboring communities with a total population of 2.4 million. At present, a first set of base year data for the three model levels have been collected, and calibration and experimental simulations are proceeding.

In the paper the simulation model as it is used in this application will be presented. The paper contains a description of the general structure of the model system and its major levels and submodels as well as a summary of the data collection and calibration problems encountered when implementing it for the urban region of Dortmund.

1. The multilevel model system

The model developed in this project is a multilevel, spatially disaggregate, recursive simulation model of regional development. It consists of three spatial levels:

(1) a macroanalytic model of the economic and demographic development of 34 labor market regions in the state of Nordrhein-Westfalen;

(2) a microanalytic model of intraregional location and migration decisions in 30 zones of the urban region of Dortmund;

(3) a microanalytic model of land use development in one or more urban districts of Dortmund.

Below, the general structure of the three-level model system and the feedbacks between the model levels are presented.

1.1. *The model levels*

The starting point of the simulation is the process of *economic change.* It is entered into the model exogenously in the form of alternative scenarios of the future development of employment and production by industry sector in the state of Nordrhein-Westfalen (cf. Rojahn, 1981). Together with additional assumptions about the demographic, social, and technological development, these scenarios are the framework for the simulation of spatial development, i.e. the simulation model shows their likely impacts on urban spatial structure.

In the normal case economic change by its requirements in terms of land, infrastructure, and housing determines the spatial development of a region, and not conversely. However, where a region fails to adapt its spatial structure to such requirements, it will slow down in its economic development and eventually lose jobs or even population to other more favored regions.

To capture such shifts of economic activity between regions it is necessary to model the *locational competition* between regions for jobs and people. This is done on the *first* spatial level of the three-level model system, the Nordrhein-Westfalen or «*Regional*» Model. Spatial units of the Regional Model are 34 *labor market regions* of the state of Nordrhein-Westfalen, hence its name (cf. Schönebeck, Wegener, 1977). The Regional Model serves to predict the development of employment by industry and population by age, sex, and nationality in each of the 34 labor market regions as well as the migration flows between them on the basis of the exogenous scenarios of economic change referred to above and under the influence of public policies in the fields of industrial development, housing, and infrastructure.

The results of the Regional Model establish the framework for the simulation of intraregional location and migration decisions in the subsequent Dortmund Region or «*Zonal*» Model. The Zonal Model constitutes the *second* level of the three-level model hierarchy. Its study area is the urban region of Dortmund consisting of 30 *zones* within the labor market region of Dortmund and four adjacent labor market regions (cf. Sonnenschein, 1976). The purpose of the Zonal Model is to simulate intraregional location decisions of industry, residential developers, and households, the resulting migration and commuting patterns, the land use development, and the impacts of public policies in the fields of industrial development, housing, and infrastructure. The model predicts for each of the 30 zones employment by industry, population by age, sex, and nationality, households, housing, and housing occupancy by type, land use by land use category, public facilities, rents and land prices.

The *third* level of the model hierarchy is the «*District*» Model. At this level the construction activity allocated to zones in the Zonal Model is further allocated to individual *tracts* within one or more zones or urban *districts*. Any zone or combination of zones could be included in the District Model, but data collection for the District Model has been limited to the 12 urban districts of Dortmund.

Fig. 1 is a schematic representation of the three model levels, the Regional Model, the Zonal Model, and the District Model and their major model sectors and interrelationships. The Regional Model and the Zonal Model both comprise the sectors employment, population, housing, and infrastructure, but the Zonal Model also includes industrial and commercial buildings, and land use. The District Model presently deals only with physical elements of urban structure, i.e. residential and nonresidential buildings and land use. On all three model levels the transportation submodel so far has not been implemented.

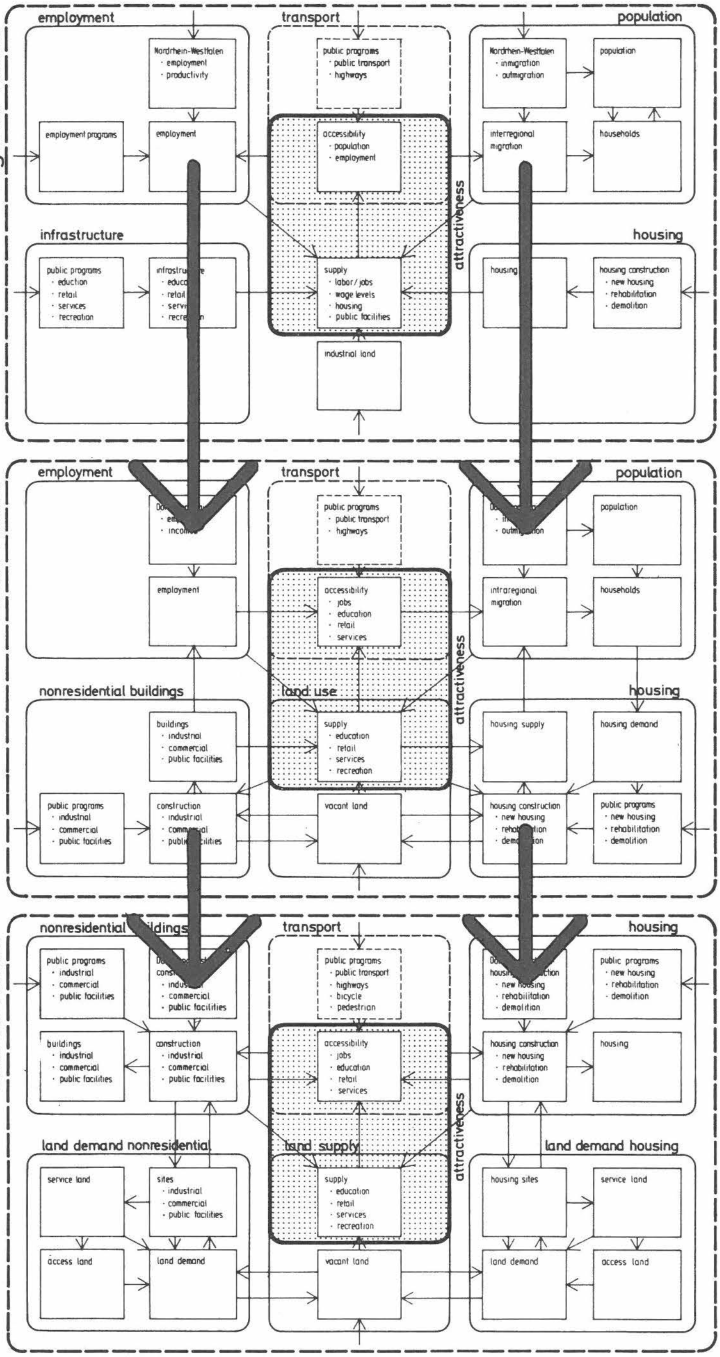


Figure 1 The three levels of the model

A central role in all three models plays the concept of *attractiveness.* The attractiveness of a region, a zone, or an object in general, is a weighted aggregate of component attributes of the object as seen and evaluated by relevant types of users, and as such it strongly influences all decisions of the model actors. In the Regional Model, the attractiveness of a region as a location for industry depends on factors such as labor supply, availability of industrial land, and the quality of housing and of the business-serving infrastructure; the attractiveness of a region for migration is mostly affected by job opportunities, wage levels, and the supply of housing and household-serving infrastructure. In the Zonal Model, the attractiveness of a zone as a location as seen by the industrial manager or residential developer is composed of attributes indicating amenities such as supply of public facilities, accessibility, or neighborhood quality, while the attractiveness of a site as seen by the developer, or of an apartment or a house as seen by a household, is an aggregate of its size, quality, and location in relation to its price. In the District Model, finally, the attractiveness of a site for a building type is determined by neighborhood amenities as well as by its zoning status and price.

The emphasis on the concept of attractiveness, and on related concepts like *utility* and *choice* clearly associates this model with a kind of economic and societal organization in which a relatively large proportion of the relevant decisions affecting urban structure are being made by *private* or semi-private actors rather than by a more or less centralized *public* authority. This is not to say that the model is geared only to a pure market economy where public intervention is low or nonexistent. Rather, the model recognizes the coexistence of market behavior and public planning interventions characteristic for the *mixed* economies of many West European countries. Accordingly, it makes a clear distinction between private decision behavior which is to be *simulated*, and public planning interventions which are *exogenously* entered. Although the volume of public intervention activity can be varied within a wide range by the model user, it can be assumed that a different type of model would be more appropriate for urban regions in countries with planned economies.

1.2. *Feedback between model levels*

The feedback between the three spatial levels of the model is established by superimposing them with the recursive *temporal* structure of the model. The model proceeds by simulation *periods* of one or two years, the end state of one period being the initial state of the next period.

Fig. 2 illustrates this superposition. The rows of this «matrix» represent the spatial levels of the model hierarchy, upside down. The columns of the matrix represent the two basic modes of operation of recursive models: the *status description* parts refer to points in time, i.e. the beginning and end of each simulation period. The *process description* parts refer to the time intervals between those points, i.e. the simulation periods. Each element of the simulation model can be located, by row and column, in this matrix.

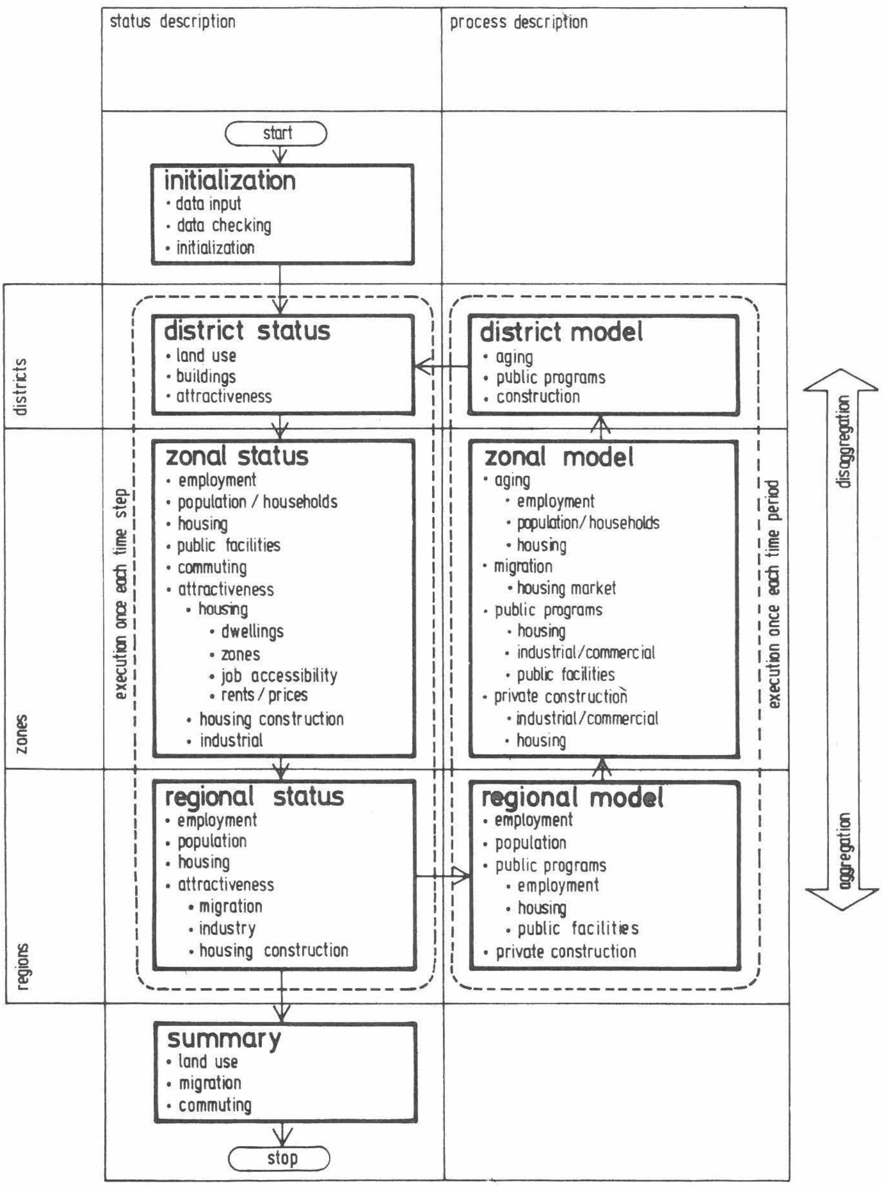


Figure 2 Feedback between model levels

The simulation begins at the symbol «start» and first passes through an initialization block. Then the recursive cycle of the model is entered. The first cycle begins with the status description of the base year, first at the lowest of the spatial levels, then by stepwise aggregation at the zonal and regional levels. At the regional level the first simulation period begins, i.e. the description of change processes between the base year and time t = 1. In the matrix, this means to step from the left to the right column. The Regional Model is executed first of the process description parts. Its results are the input to the Zonal Mode, and so forth, until eventually the results have been disaggregated down to the detail level of the District Model. That closes the first simulation period. The model again changes to the left column of the matrix and starts, with different state values, the next status description. This cycle of aggregation and disaggregation is iterated for each simulation period, until the last period has been simulated. In this case the model proceeds through a final report phase and closes down at the symbol «stop».

It must be noted, however, that this is an idealized description of the feedback processes between model levels. Due to computing time restrictions, at present the three model levels are operated only separately with data files serving as communication links between them. This means that presently only unidirectional information flows from the regional down to the zonal and district levels (as indicated by the heavy arrows in fig. 1) are in effect.

2. Model levels and submodels

In this main part of the paper the Regional Model, the Zonal Model, and the District Model, and their major submodels will be presented in more detail. The presentation follows working papers by their authors (cf. Schönebeck, 1981; Wegener, 1981; Tillmann, 1981).

2.1. *The Nordrhein-Westfalen Model (Regional Model)*

The Regional Model constitutes the first level of the three-level model system. Its purpose is to forecast employment and population in 34 labor market regions of Nordrhein-Westfalen on the basis of exogenous projections for the state and under the influence of public policies in the fields of industrial development, housing, and infrastructure.

The 34-region system exhaustively covering the state of Nordrhein-Westfalen is shown in fig. 3. The regions range in population between under 200,000 and about 1.7 million for the regions of Cologne and Düsseldorf, with the region of Dortmund (1.1 million) ranging third. Total population in all 34 regions is about 17 million.

In the *status description* part of the Regional Model the 34 regions are evaluated as industrial and residential locations. Industry is represented in the model by jobs and enterprises of 40 industrial sectors. The model population is disaggregated by age, sex, and

Figure 3 The regions of the Nordrhein-Westfalen Model (Regional Model)

nationality, and in the form of households by size and income. Labor-force participation is distinguished by sex and skill. Housing has been classified by size and quality into nine housing types. The model infrastructure includes education, retail, services, recreation, and transport facilities. From these data for each region *attractiveness indicators* for different industries and population groups are calculated. They will be discussed later.

In the *process description* part of the Regional Model the competition and exchange relations between the labor market regions are simulated as interregional migrations of labor and capital based on group-specific attractiveness differentials. The process description part consists of four submodels: the *economic* submodel, the *demographic* submodel, the *public programs* submodel, and the *private construction* submodel. For space limitations, only the first two of them will be described here. Of the remaining two, the *public programs* submodel serves to process user-specified public programs in the fields of industrial development, housing, and educational, recreational, and transport infrastructure, while in the *private construction* submodel changes of the housing stock of the regions by private developers are simulated (cf. Schönebeck, 1981).

2.1.1. The economic submodel

In this submodel employment by 40 industry sectors is predicted for each of the 34 labor market regions in accordance with exogenous projections for sectoral employment in Nordrhein-Westfalen.

Given this sectoral and spatial disaggregation, data restrictions made it impossible to follow the mainstream of multiregional economic modeling, i.e. to adopt an econometric or a multiregional input-output model (cf. Bolton, 1980). In fact, the primary available source of information were employment data by industry and region for several points in time. This made it necessary to adopt a model which makes the best use of the trend information contained in such data, but at the same time takes account of developments which are likely to modify these trends.

This is accomplished by combining elements of conventional *shift-and-share* analysis with a causal model relating *changes* of the locational component of the shift-and-share analysis to *changes* in the perceived locational attractiveness of the regions. The following description illustrates this:

at first for each industrial sector s and region i two *fictitious* employment levels for time t + 1 are calculated. The one, E''_{si}, might be expected if all regions continued to grow or decline at the same rate as in the previous period, provided this was possible under the sectoral forecast for Nordrhein-Westfalen, i.e. if the *pattern* of, or relations between, regional rates of change stayed the same. The other, E''_{si}, is the employment level to be expected if all regions developed according to the sectoral forecast for the whole state, i.e. irrespective of their previous development:

$$E'_{si}(t+1) = \frac{\dfrac{E_{si}(t)}{E_{si}(t-1)} E_{si}(t)}{\sum\limits_{i} \dfrac{E_{si}(t)}{E_{si}(t-1)} E_{si}(t)} E^{*}_{s}(t+1) \qquad (1)$$

$$E''_{si}(t+1) = \frac{E^*_s(t+1)}{E^*_s(t)} E_{si}(t) \qquad (2)$$

where $E^*_s(t)$ is total employment of sector s at time t. Obviously, neither of the two projections is realistic as there *is* variation between regional rates of development, and this variation will *change* as winning regions may turn to losers, and vice versa.

The model proceeds by dividing the «trend» projection E'_{si} with the help of the «ahistoric» projection E''_{si} into two multiplicative components (Birg, 1975):

$$q_{si} = \frac{E''_{si}(t+1)}{E_{si}(t)} \qquad (3)$$

$$r_{si} = \frac{E'_{si}(t+1)}{E''_{si}(t+1)} \qquad (4)$$

where q_{si} is the *structural factor* or that proportion of the rate of change of employment which is attributable to normal sectoral development, and r_{si} is the *locational factor* or residual attributable to location, i.e. to the perceived attractiveness of the region as a location for that industry. If that attractiveness changes, the locational factor should change, too.

At this point the attractiveness indicators calculated in the status description part of the model are used for the first time, therefore a brief discussion of the formal properties of the concept of attractiveness used throughout the model levels is appropriate. The utility model used is the additive multiattribute utility theory (MAUT) model

$$A_{ni} = \frac{1}{\sum_m w_{mn}} \sum_m w_{mn} v_{mn}(a_{mi}) \qquad (5)$$

where A_{ni} is the attractiveness of evaluation object i for activity n, a_{mi} is the m-th attribute of that evaluation object, and w_{mn} and v_{mn} are importance weights and value or utility functions, respectively, of that attribute as seen by actor type n. The v_{mn} can be any function relating values of attributes to a standardized utility scale of, say, between 0 and 100.

In the Regional Model the actor types are industries by sector or, in the demographic submodel, migrants, and the evaluation objects are the regions. Regional attributes can be indicators for amenities supplied in the regions themselves or accessibility measures:

$$a_{mi} = f_m(b_{ki}) \qquad (6a)$$

$$a_{mi} = \sum_j \frac{f_m(b_{kj}) \exp(-\beta c_{ij})}{\sum_j f_m(b_{kj}) \exp(-\beta c_{ij})} c_{ij} \tag{6b}$$

where $f_m(b_{ki})$ is a generating function specifying how to calculate a_{mi} from the k-th variable b of region i, and c_{ij} is an indicator for travel time or cost between regions i and j.

In the Regional Model the attractiveness of a region as a location for enterprises of a given industrial sector is represented by attributes such as:

– the *labor market situation* expressed by supply and demand on four labor submarkets distinguished by sex and skill, taking account of actual and potential labor-force participation and commuting,

– the availability of *industrial land*,

– the availability of *financial aid*,

– the quality of the *housing* stock,

– the quality of the business-serving *infrastructure*,

– the existing *industrial mix.*

Obviously, for each industrial sector a different set of attractiveness functions using different attributes, weights, and utility functions is required as enterprises of different sectors have different locational preferences.

The resulting attractiveness A_{si} of region i for industry s is compared with a fictitious attractiveness A'_{si} which the region might have if everything in it had developed in line with the trend of the previous period. If there *is* a difference, it is used to estimate a corresponding change of the *locational factor*

$$r'_{si} = r_{si} + f_s(A_{si} - A'_{si}) \tag{7}$$

where r'_{si} is the modified locational factor and f_s is a (linear) function of the deviation of regional attractiveness from the attractiveness trend. It should be noted that both, locational factors and attractiveness indicators, are smoothed by a first order exponential delay (Schönebeck, Wegener, 1977) in order to avoid excessive oscillation. Then

$$E_{si}(t+1) = q_{si} r'_{si} E_{si}(t) \tag{8}$$

is the final employment prediction equation.

2.1.2. The demographic submodel

In the demographic submodel the population of the labor market regions is aged by one period, and the interregional migrations are predicted. Both steps are performed independently in order to take account of dynamic changes of reproduction and migration behavior.

Aging of the population is performed twice per simulation period, half a period each, before and after migration. The aging model uses conventional cohort survival techniques adapted to five-year age groups, and is based on time-invariant life tables, but dynamic, age-specific, and regionalized fertility estimates. In addition, in each period a proportion of the foreign population is moved to the native state by naturalization.

The *migration* model consists of a set of age-specific, deterministic, production-constrained spatial interaction models (cf. Gatzweiler, 1975). Five age groups of migrants are considered: 16-20, 21-35, 36-50, 51-65, and over 65 years, a sixth age group is reserved for children of age 0-15 migrating with their parents. In addition, migrations are distinguished by nationality.

For each of the 5 × 2 adult age groups the following steps are performed: *First*, total interregional migration volume is estimated on the basis of the volume of the previous period and the number of inmigrations into and outmigrations out of Nordrhein-Westfalen, the latter being exogenously specified. *Second*, the number of migrations out of each region is estimated as a proportion of total interregional migration:

$$M^{o}_{gni}(t,t+1) = \frac{[P_{gni}(t) - H_{gni}(t)][100 - A_{gni}(t)]}{\sum_i [P_{gni}(t) - H_{gni}(t)][100 - A_{gni}(t)]} M^{*}_{gn}(t,t+1) . \qquad (9)$$

In this equation $M^{o}_{gni}(t,t+1)$ are the migrations of age group g and nationality n originating from region i between t and t+1. $P_{gni}(t)$ and $H_{gni}(t)$ are population and home-owning population, respectively, i.e. only population not living in owner-occupied homes are considered potential migrants. $A_{gni}(t)$ is the attractiveness of the region for migration, which will be discussed in the next paragraph. $M^{*}_{gn}(t,t+1)$ are total interregional migrations of age group g and nationality n between t and t+1. *Third*, the migration origins thus established are distributed by the following production-constrained model:

$$M_{gnij}(t,t+1) = \frac{D_j(t)\, A_{gnj}(t) \exp(-\beta c_{ij})}{\sum_j D_j(t)\, A_{gnj}(t) \exp(-\beta c_{ij})} M^{o}_{gni}(t,t+1) . \qquad (10)$$

Here, $M_{gnij}(t,t+1)$ are migrations between regions i and j. The $D_j(t)$ are attractor variables of the target regions such as jobs or educational facilities, or weighted aggregates thereof. The A_{gnj} again are attractiveness indicators.

It will have been noted that the attractiveness indicators are used in (10) as *pull* or attraction variables and in (9), with a negative sign, as *push* or deterrence variables. In both cases the same weighted aggregate of regional attributes is used which may be decomposed into two subsets:

– attributes expressing the *labor market situation*, such as job opportunities and wage levels,

– attributes expressing the *housing situation*, such as housing supply, household-serving infrastructure, and environmental quality.

In contrast to other attractiveness indicators used in the model, weighting between these two subsets is done *dynamically* in response to statewide labor market conditions to take account of the fact that at times of high unemployment job considerations become of primary importance for migration decisions.

In a final step the number of children migrating with their parents is estimated for parent age groups, and migrations into and out of Nordrhein-Westfalen are assigned to regions in proportion to interregional migration.

2.2. *The Dortmund Region Model (Zonal Model)*

The Zonal Model establishes the second level of the three-level model hierarchy. Its study area is the *urban region* of Dortmund consisting of Dortmund itself with its 12 urban districts and ten neighboring communities within the labor market region of Dortmund, plus eight zones in four adjacent labor market regions. This 30-zone system and its superposition with the system of labor market regions of the Regional Model is shown in fig. 4. The 12 urban districts of Dortmund are relatively homogenous in size, ranging in population between 40,000 and 60,000, while the remaining zones vary considerably in population between about 15,000 and over 400,000 (Bochum). The whole urban region has a population of about 2.4 million.

The purpose of the Zonal Model is to simulate intraregional location decisions of industry, residential developers, and households, the resulting migration and commuting patterns, the land use development, and the impacts of public policies in the fields of industrial development, housing, and infrastructure.

In the *status description* part of the model the 30 zones are evaluated in terms of employment, population, housing, industrial and commercial buildings, and public facilities. Employment and industrial and commercial buildings are disaggregated by the same 40 industry sectors as in the Regional Model. Population is disaggregated by age, sex, and nationality. In addition, the population of each zone is represented as a distribution of households classified by

– nationality (native, foreign),
– age of head (16-29, 30-59, 60+ years),
– income (low, medium, high, very high),
– size (1, 2, 3, 4, 5+ persons).

Similarly, housing of each zone is represented as a distribution of dwellings classified by

- type of building (single-family, multi-family),
- tenure (owner-occupied, rented, public),
- quality (very low, low, medium, high),
- size (1, 2, 3, 4, 5+ rooms).

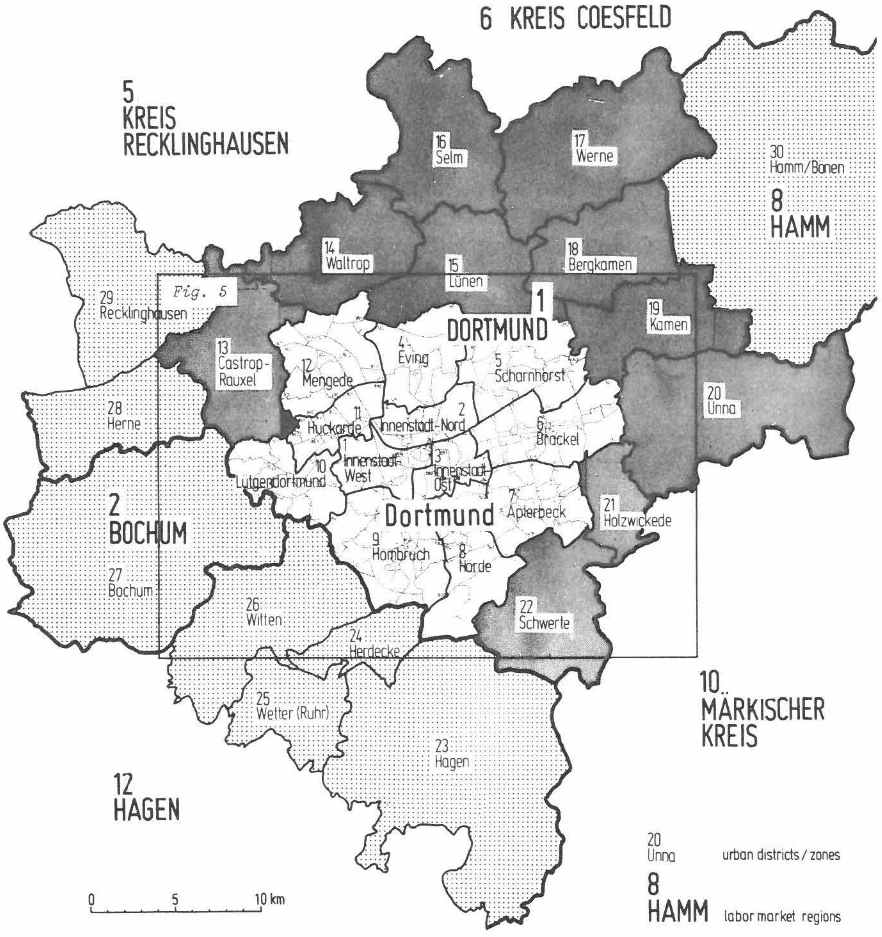


Figure 4 The zones of the Dortmund Region Model (Zonal Model)

All changes of population and housing during the simulation are computed for these 120 household types and 120 housing types. However, these household and housing types are collapsed to M household and K housing types for use in the *occupancy matrix*, with M and K not exceeding 30.

The *occupancy matrix* **R** of a zone is an M × K matrix representing the association of households with housing in the zone. Each element R_{mk} of the matrix contains the number of households of type m occupying a dwelling of type k, the total matrix contains all households occupying a dwelling or all dwellings occupied by a household. In addition, there are two M × 1 vectors, $\mathbf{H}^s$ and $\mathbf{H}^w$, containing subtenant households and households *without* a dwelling, respectively, and two 1 × K vectors, $\mathbf{D}^n$ and $\mathbf{D}^v$, containing new and vacant dwellings, respectively.

Infrastructure is represented in the model by various facilities from the fields of health, welfare, education, recreation, and transport.

As in the Regional Model, these data are used to calculate *attractiveness* indicators for different actor groups. Six sets of attractiveness indicators are required for the Zonal Model: the attractiveness of *zones* and *land use categories* for different industries and residential developers, and the attractiveness of *zones* and *housing types* for different household types.

In the *process description* part of the Zonal Model location and migration decisions of enterprises, residential developers, and households are simulated. This implies the simulation of major changes of employment, population, housing, and infrastructure facilities. Such changes can be caused by *time (aging)*, *migration*, *public programs*, or *private construction.* In the Zonal Model these four kinds of change are treated in four separate submodels. For space reasons, only three of them can be described here. The *public programs* submodel not described processes a large variety of public programs in the fields of employment, housing, and health, welfare, education, recreation, and transport facilities.

2.2.1. The aging submodel

In the first, the *aging* submodel, all changes of the model variables are computed which are assumed to result from biological, technological, or long-term socioeconomic trends originating outside of the model, i.e. which in the model are merely *time*-dependent. These changes are effected in the model by probabilistic aging or updating models. Presently there are three such models for employment, population, and households/housing.

For updating zonal *employment* at this point simply the regionwide rates of change provided by the Regional Model for each industrial sector are projected on zonal employment. However, it is checked if the resulting employment levels can be accommodated into the existing industrial and commercial buildings, and if they cannot, they are reduced accordingly. The resulting employment levels are still tentative and will be modified in the *public programs* and *private construction* submodels.

The *population* projection model is identical to that of the Regional Model predicting zonal population by age, sex, and nationality on the basis of time-invariant life tables and naturalization rates, and dynamic, age-specific, and spatially disaggregate fertility estimates.

For *households* time-dependent changes include demographic changes of household status in the life cycle of the household such as birth, aging, death, marriage, and divorce, and all new or dissolved households resulting from these changes, as well as change of nationality or income. On the *housing* side they include deterioration and certain types of rehabilitation and demolition. However, all changes of housing occupancy connected with migration decisions are left to the subsequent *migration* submodel, i.e. the aging model *ages* all households and dwellings by one simulation period *without* moving them relative to each other. This is accomplished by a Markov model with dynamic transition rates.

A transition rate is defined as the probability that a household or dwelling of a certain type changes to another type during the simulation period. The transition rates are computed as follows: the time-dependent changes to be simulated are interpreted as *events* occurring to a household or dwelling with a certain probability in a unit of time. These *basic event probabilities* and their expected future development are exogenously determined. Fifteen basic event probabilities have been identified for each of the three household age groups:

1 change of nationality
2 aging
3 marriage
4 birth, native
5 birth, foreign
6 relative joins household
7 death
8 death of child
9 marriage of child
10 new household of child
11 divorce
12 rise of income
13 decrease of income
14 retirement
15 new job

and three for the four housing quality groups:

1 deterioration
2 rehabilitation
3 demolition.

Not all household events occur to every household. Some are applicable only to singles, some only to families, some only to adults, sone only to children. Some household events are followed by housing events, and vice versa: where a household dissolves, a dwelling is vacated, and where a nonvacant dwelling is demolished, a household is left without dwelling. The housing events contain only those changes of the housing stock which can be expected to occur under normal conditions in any housing area, i.e. a normal rate of deterioration, maintenance, rehabilitation, and demolition. More rehabilitation and demolition may occur later in the *private construction* submodel: rehabilitation as a response of housing investors to the demand situation observed on the housing market, demolition where housing has to make way for industrial or commercial land uses. In addition, rehabilitation and demolition may occur in the course of public construction programs in the *public programs* submodel.

The basic event probabilities are then aggregated to transition rates **h** for households and **d** for dwellings using the disaggregate (120-type) household and housing distributions of each zone. Most events are independent of each other and can be aggregated multiplicatively; but some exclude others, i.e. are the complement of each other. The matrices **h** and **d** are of dimensions $M \times M$ and $K \times K$, respectively, where the rows indicate the source state and the columns the target state. Multiplication of **h** and **d** with the occupancy matrix **R** yields the occupancy matrix aged by one simulation period:

$$\mathbf{R}(t+1) = \mathbf{h}'(t,t+1)\,\mathbf{R}(t)\,\mathbf{d}(t,t+1) \tag{11}$$

where $\mathbf{h}'$ is the transpose of **h**. This implies the assumption that all households share the same transition rates, no matter in which dwelling they live, and vice versa.

Special provisions are necessary for events which create new households without a dwelling or new vacant dwellings. New households without a dwelling may be generated by marriage of child, new household of child, or divorce:

$$\mathbf{H}^n(t,t+1) = \mathbf{h}^n(t,t+1)\,\mathbf{H}^*(t) \tag{12}$$

where $\mathbf{h}^n(t,t+1)$ is an $M \times M$ matrix containing current household formation probabilities. An element $h^n_{mp}(t,t+1)$ of this matrix is defined as the probability that a new household of type m is produced by a household of type p during the simulation period. $\mathbf{H}^*(t)$ is the $M \times 1$ vector of total households. Another way that a household without a dwelling may be generated is by demolition of a dwelling:

$$\mathbf{H}^d(t,t+1) = \mathbf{R}(t)\,\mathbf{d}^d(t,t+1) \tag{13}$$

where $\mathbf{d}^d(t,t+1)$ is a $K \times 1$ vector of demolition rates of housing types. Similarly, new vacant dwellings may be generated by dissolution of households:

$$\mathbf{D}^d(t,t+1) = \mathbf{h}^d(t,t+1)\,\mathbf{R}(t) \qquad (14)$$

where $\mathbf{h}^d(t,t+1)$ is a $1 \times M$ vector of dissolution rates of households aggregated from basic events like marriage, relative joins household, and death. Of course, new dwellings may also result from housing construction, but that is effected in the *public programs* and *private construction* submodels.

In addition, it is necessary to age households and dwellings outside of the **R** matrix as also households without dwellings get older, and vacant dwellings deteriorate or may be rehabilitated or be torn down:

$$\mathbf{H}(t+1) = \mathbf{h}'(t,t+1)\ [\mathbf{H}^s(t) + \mathbf{H}^w(t) + \mathbf{H}^d(t,t+1) + \mathbf{H}^n(t,t+1)] \qquad (15)$$

$$\mathbf{D}(t+1) = [\mathbf{D}^v(t) + \mathbf{D}^d(t,t+1) + \mathbf{D}^n(t-1,t)]\ \mathbf{d}(t,t+1) \qquad (16)$$

where $\mathbf{D}^n(t-1,t)$ is the $1 \times K$ vector of dwellings newly constructed in the previous period. In equations (15) and (16) all households without a dwelling and all vacant dwellings of a zone are consolidated into two vectors **H** and **D** for use in the subsequent housing market model.

2.2.2. The migration submodel

In the second, the *migration* submodel intraregional migration decisions of households are simulated. In contrast to interregional migrations, intraregional migrations are largely determined by housing considerations. Consequently, the intraregional migration model was designed as a housing market model.

The simulation technique selected for this model is the Monte Carlo technique. Stochastic micro simulation techniques have only recently been introduced into housing market modeling. Several models use probabilistic micro approaches to model the aging process of households and housing (e.g. Kain *et al.*, 1976; Schacht, 1976). However, with the exception of one early example (Azcarate, 1970), no stochastic models of household decision behavior have been reported.

When the housing market simulation is entered, the following situation exists: all households and dwellings of all zones have been aged by one simulation period, i.e. now have the time label of the end of the current simulation period. However, no household has yet moved to another dwelling. That is to say: all households have proceeded in their life cycle - they have become older, children may have been born, the family income may have increased -, but their dwellings are still the same or even have deteriorated. Moreover, the expectations of the households with respect to size, quality, and location of housing generally will have increased. It may be assumed that many households which were quite satisfied with their housing situation at the end of the last simulation period now are dissatisfied with it and are willing to improve it.

These households are the potential movers of the current market period. They are contained in the occupancy matrix **R** of each zone. Besides, there are households without dwellings contained in the vector **H** and vacant dwellings contained in the vector **D** of each zone. In addition, there are two $M \times 1$ vectors of households specified by the Regional Model: the vector $\mathbf{H}^e$ containing households migrating into the region from elsewhere during the simulation period, and the vector $\mathbf{H}^o$ containing households migrating out of the region. Inmigrant households are treated just like households without dwelling, except that they do not come from a particular zone. Outmigrant households are of interest because they vacate a dwelling.

Unlike in the aging submodel, now the information of all 30 zones has to be available simultaneously. Therefore, by the additional zonal dimension, the matrix **R** becomes three-dimensional, and the vectors **H** and **D** become two-dimensional matrices. The matrices **R**, **H**, and **D**, plus the vectors $\mathbf{H}^e$ and $\mathbf{H}^o$ are a complete representation of households and housing at the outset of the market simulation.

Of these the matrix **H** and the vector $\mathbf{H}^e$ clearly represent housing demand, and the matrix **D** and the vector $\mathbf{H}^o$ clearly represent the supply side. The matrix **R** represents some of both because of the linkage between housing supply and housing demand by vacant dwellings being put on the market with each move. But which of the households in **R** will actually move during the market period is not known at this moment.

The decision behavior of the model actors in the housing market model is controlled by attractiveness indicators which are called *housing satisfaction* here. The satisfaction of a household with its housing situation is represented in the model by a multidimensional attractiveness function containing the dimensions *housing size and quality*, *neighborhood quality*, *location*, and *housing cost.* Two of these four dimensions are themselves multiattribute:

– *housing size and quality* is composed of the attributes defining a housing type: type of building, tenure, quality, size;

– *neighborhood quality* is composed of attributes selected or aggregated from zonal variables from the fields of population, employment, buildings, public facilities, transportation, and land use as well as of accessibility measures indicating the location of the zone to the work places and to retail, education, and recreation facilities in other zones.

The remaining two dimensions have only one attribute: the *location* dimension is represented by the attribute «job accessibility», while the only attribute of *housing cost* is rent or housing price in relation to income.

For use in the housing market simulation the four dimensions of housing satisfaction are stored in two matrices: for each combination of household type m, dwelling type k, and zone i, i.e. for each element of the three-dimensional occupancy matrix **R**, an *index of housing satisfaction* U_{mki} is calculated as a weighted aggregate of the four dimensions. Obviously, in this index only a general measure of job accessibility like that in (6b) can be included. Therefore an additional location measure is calculated for each pair of zones:

$$W_{ii'} = \sum_j \frac{T_{ij} T_{i'j}}{\sum_j T_{ij} T_{i'j}} \; v(c_{i'j}) \tag{17}$$

where T_{ij} are home-to-job trips from i to j, $T_{i'j}$ are job-to-home trips from j to i′, and $v(c_{i'j})$ is a utility function of job accessibility. That is, $W_{ii'}$ expresses the attractiveness of zone i′ as a new housing location with respect to job accessibility for a household now living in zone i whose head has a job in zone j. The measure $W_{ii'}$ is called the *migration distance* between i and i′.

With the matrices **R**, **H**, **D**, **U**, and **W**, and the vectors $\mathbf{H}^e$ and $\mathbf{H}^o$ all necessary information is available to enter the housing market simulation, i.e. the simulation of the market clearing process.

The Monte Carlo technique used is based on the notion that the total market process can be sufficiently approximated by simulating a representative sample of individual market transactions. To achieve this, the model consists of a sequence of random selection operations by which hypothetical market transactions are generated. The random selection process is controlled by probability distributions which insure that only likely transactions are selected.

The basic unit of the simulation thus is the *market transaction*. A market transaction is any successfully completed operation by which a migration occurs, i.e. a household moves into or out of a dwelling or both. There are two ways to start a market transaction: a household decides to look for a dwelling (« *dwelling wanted* »), or a landlord decides to offer a dwelling (« *dwelling for rent or sale* »). In either case the transaction may result in different kinds of migration: the household may leave the region (« *outmigration* ») or enter it (« *inmigration* »), or currently be without a dwelling (« *new household or forced move* »), or occupying one (« *move* »). For the landlord offering a dwelling only the last three migration types are of interest.

The model starts by selecting a transaction type and a migration type. It is assumed that « dwelling wanted » and « dwelling for rent or sale » are equally likely to occur. The migration type is selected in proportion to the number of migrations to be completed of each type, i.e the totals of $\mathbf{H}^o$, $\mathbf{H}^e$, and **H** for the first three migration types, respectively.

For the fourth or «move» type a tentative estimate of the number of moves as a portion of the matrix **R** must be provided.

Once the transaction type and the migration type have been determined, the remaining parameters of the transaction are selected. A transaction has been completely defined if the following six parameters are known:

m household type
k old housing type
i old zone
j zone of job
k′ new housing type
i′ new zone.

A move, for instance, is a migration of a household of type m which occupies a dwelling of type k in zone i and whose head works in zone j, into a dwelling of type k′ in zone i′. Not all six parameters are required for all migration types: obviously, no k can be specified for households without dwelling, nor can k and i for inmigrant households, but it is assumed that inmigrant households have a job in j already. Of outmigrant households only m, k, and i are of interest.

In each step one additional parameter is determined, until the transaction has been completely defined. The following example illustrates this: in the case of a household considering a move («dwelling wanted» and «move») first the household by type, zone, and dwelling type is selected with

$$p(k|mi) = \frac{R_{mki}(100-U_{mki})^{\alpha}}{\sum_k R_{mki}(100-U_{mki})^{\alpha}} \tag{18}$$

being the probability of dwelling type k to be selected if household type m and zone i are already known, which is to say that households which are dissatisfied with their housing situation are selected more often than others. In the next two steps it is asked in which zone j the head of the household might have his job and how this may restrict the choice of a new housing zone. With the help of the *migration distance* defined in (17) these two selection steps can be collapsed into one with

$$p(i'|mki) = \frac{\sum_{k'} D_{k'i'} W_{ii'}^{\beta}}{\sum_{j'}\sum_{k'} D_{k'i'} W_{ii'}^{\beta}} \tag{19}$$

being the probability of zone i′ to be selected as a new housing zone where m, k, and i are given and zone k assumed to be the work place zone of the household head. In the final selection step the household attempts to find a dwelling in zone i′ with

$$p(k'|mkii') = \frac{D_{k'i'} U_{mk'i'}^{\gamma}}{\sum_{k'} D_{k'i'} U_{mk'i'}^{\gamma}} \quad (20)$$

being the probability of dwelling type k′ to be selected if all other parameters are given.

Once the transaction has been completely defined, the migration decision is made. This is not a question for outmigrant households, they do migrate. All other households compare their present housing situation with the situation they would gain if they accepted the transaction. It is assumed that they accept it if they can significantly improve their housing situation. The definition of what is considered significant has to be determined by calibration.

If there *is* a significant improvement, the household accepts the new dwelling. In this case all necessary changes in **R**, **H**, $\mathbf{H}^e$, $\mathbf{H}^o$, and **D** are immediately performed. Dwellings vacated with a move or an outmigration reappear in the matrix **D** and are thus again released to the market.

If there is no improvement involved in the move, the household declines. It makes another try to find a dwelling, and with each attempt it accepts a lesser improvement. After a number of unsuccessful attempts it abandons the idea of a move. The landlord tries to find another household, but he does not reduce the rent during the market period. If a dwelling type in a zone has been refused by all household types, it is taken out of the market for this period.

After successful or unsuccessful completion of a transaction the next transaction is selected. The market process comes to an end when there are no more households considering a move. It is assumed that this is the case when a certain number of transactions has been rejected. This number has to be determined by calibration to match the number of migrations produced by the model with the number of migrations observed in the region.

The results of the housing market simulation serve to calculate migration flows and migration-induced changes of the age and household distributions of the zones. In addition, rents and housing prices are adjusted in response to the demand observed on the market for use in the next simulation period.

2.2.3. The private construction submodel

In the *private construction* submodel investment and location decisions of the great number of private developers are modeled, i.e. of enterprises which erect new industrial or commercial buildings, and of residential developers who build apartments and houses for sale or for rent or for their own use. Thus the submodel is a model of the regional *land and construction market.*

The model treats each industry sector and housing type as a separate submarket and processes them sequentially, i.e. in the model there is no competitive bidding for a unit of land by different kinds of building use, although this is a common occurrence in reality. The problem is taken account of by preordering the building uses by decreasing profitability and processing them in that order.

For each submarket the following three steps are performed: first, the volume of construction *demand* of the particular building use in the current period is estimated. Second, the *capacity*, i.e. zoned vacant land, of each zone for that building use is determined. Third, the estimated volume of construction is allocated to the vacant land of the zones as a function of their *attractiveness.*

The *demand* for new industrial or commercial buildings of a particular industry sector is estimated by comparing the number of jobs in that sector accommodated in existing buildings in the whole region with the employment projected by the Regional Model for the end of the period. Of the balance, some jobs may move into buildings vacated by other industries, but for the rest new floor space must be provided. The demand for different housing types is estimated in response to the housing demand observed on the housing market with the effect that no more housing will be built of types where there are many vacant units already.

The *capacity* of each zone for a particular building use is determined by searching the zone for vacant land suited for that building use. For this purpose the model contains a disaggregate land use inventory comparable to a *zoning plan* including such information as existing use, zoned land use category, maximum density, and time restrictions. The model searches through the land use inventory and, taking account of floor space requirements per work place or dwelling unit, calculates the number of work places or dwelling units that *could* be accommodated in each zone. In addition, it is estimated where, in the case of high demand, additional building space *could* be procured by demolition or change of use of existing buildings.

If demand and capacity for a particular building type are known, the actual allocation can be performed. It is assumed that, where capacity or supply of land exceeds demand, utilization of capacity varies with its *attractiveness* for the enterprise or developer. The attractiveness of a unit of land as a location consists of zonal as well as of site attributes aggregated into three subsets:

- *neighborhood quality* composed of zonal attributes and accessibilities,
- *suitability of land use category* including such attributes as maximum density and environmental and other restrictions,
- *land price* in relation to expected profit.

Needless to say that the attractiveness functions, i.e. the attributes, weights, and utility functions, are different for different building types.

The allocation model has the following form - note that construction is allocated not to zones, but to land use categories within zones:

$$C_{nli}(t,t+1) = \frac{K_{nli}A_{nli}(t)^{\alpha}}{\sum_{i}\sum_{l} K_{nli}A_{nli}(t)^{\alpha}} C^{*}_{n}(t,t+1) \quad (21)$$

where K_{nli} is the capacity, $A_{nli}(t)$ is the attractiveness, and $C_{nli}(t, t+1)$ is the amount of construction of building type n allocated to land use category l in zone i of total demand $C^{*}_{n}(t, t+1)$ for that building type in the period.

The allocation model is iterated to account for new demand caused by demolition of existing buildings. After the last iteration land prices are adjusted in response to observed demand for use in the next period.

2.3. *The Urban District Model (District Model)*

The third level of the three-level model hierarchy is established by the District Model. Its study area is confined to one or any group of adjacent *urban districts* of Dortmund, which are identical to zones 1-12 of the Zonal Model. Spatial units of the District Model are statistical subdistricts or *tracts*, the next higher level above the statistical block. The tracts range in population between a few hundred and as much as 20,000, but the majority of them lie between 2,000 and 5,000. There are between 9 and 28 tracts per district, and 171 in the whole city. The system of tracts and districts in Dortmund is shown in fig. 5.

The purpose of the District Model is to allocate the construction activity simulated in the Zonal Model to individual tracts within the zones or urban districts. Therefore the District Model presently deals only with physical elements of urban structure, i.e. residential and nonresidential buildings and land use, but has no explicit employment and population sectors.

The model structure of the District Model resembles that of the Zonal Model. It has an *aging* or updating, a *public programs*, and a *construction* submodel, but no migration or housing market submodel. The aging submodel serves to update the building stock by one simulation period. It is a probabilistic model like its counterpart in the Zonal Model, but is much simpler as no association of households and housing is necessary. The public programs submodel allows to enter public programs in the fields of housing and public facilities also on the neighborhood level. The construction submodel is the core of the District Model. It will be described below.

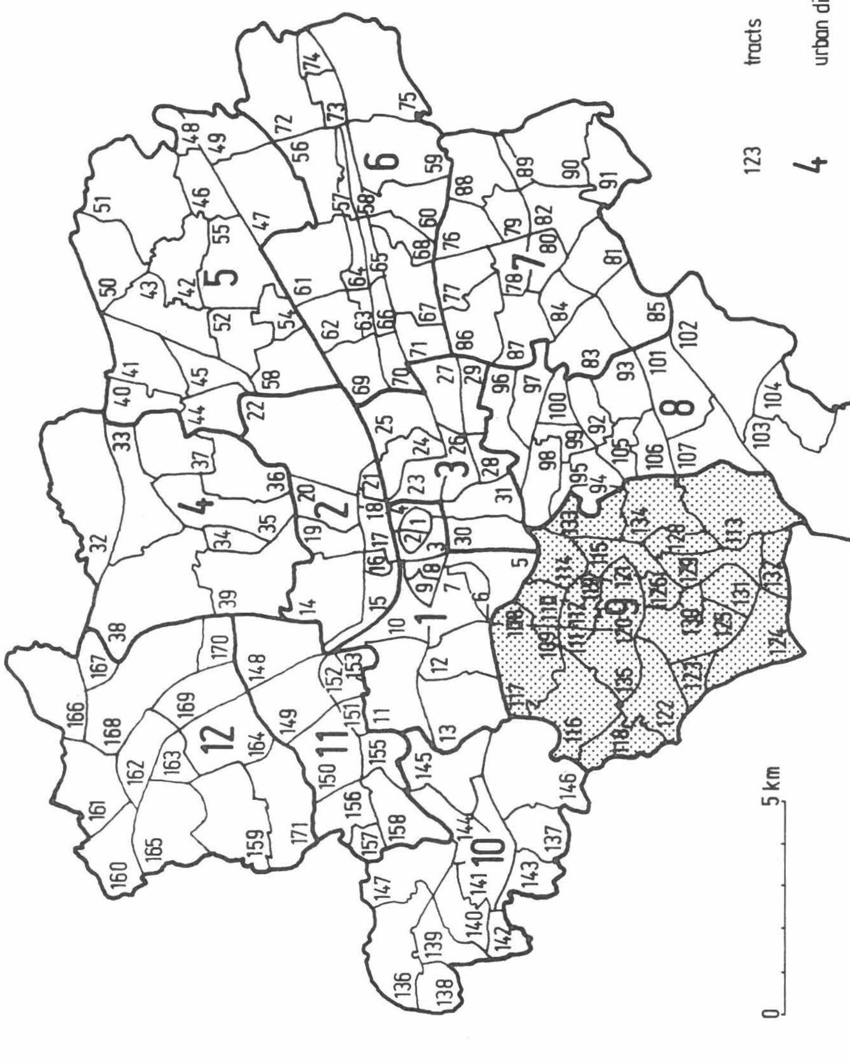


Figure 5 Tracts and districts in Dortmund

In the *construction* submodel public and private construction activity allocated to the zones in the Zonal Model - industrial and commercial buildings by industrial sector, housing by housing type, and public facilities by type - are allocated to individual tracts within the zone. This is accomplished in a similar way as in the Zonal Model, with a separate allocation model for each building type, whereby the problem of competitive bidding for land is solved by preordering the building types before allocation.

The allocation model itself is of the same type as in the Zonal Model (21). However, now the *zoning plan* governing the allocation capacity of tracts is much more detailed including, for instance, such additional information as maximum number of floors of buildings. In addition, now indivisibilities of construction projects are taken into account by allowing only complete buildings B_n of minimum size b_n to be built:

$$B_{nli}(t,t+1) = \text{int}\left[\frac{1}{b_{ni}} \frac{K_{nli}A_{nli}(t)^{\alpha}}{\sum_i \sum_l K_{nli}A_{nli}(t)^{\alpha}} C_n^z(t,t+1)\right] \qquad (22)$$

where int(x) means the largest integer less or equal x. The subscript i now represents tracts instead of zones, while $C_n^z(t, t+1)$ is total construction volume of building type n in the zone. The integrity condition requires iteration to satisfy the constraint

$$C_n^z(t,t+1) = \sum_i \sum_l B_{nli}(t,t+1)\, b_{ni} \,. \qquad (23)$$

In analogy to the Zonal Model, the attractiveness A_{nli} in (22) is an aggregate of tract as well as of site attributes assembled into three subsets: *neighborhood quality* composed of tract attributes and accessibilities, *suitability of land use category* composed of zoning and other restrictions, and *land price* in relation to expected profit.

In a final step for each building the number of floors and the distribution of building uses such as dwellings, shops, offices, etc. on each floor are estimated in accordance with the zoning plan.

3. Model data and calibration

While the collection of a first set of base year data for the three model levels has been completed, calibration and test of the model are still under way. Therefore, only a brief summary of the data and calibration problems encountered so far will be given here.

Base year of the simulation presently is 1970. As ten two-year periods can be simulated, the forecasting horizon presently is 1990. In this case the first four or five simulation periods serve as calibration or test periods. However, it is planned to move the base year to the present and thus extend the forecasting horizon up to the year 2000.

Most *base year data* for the *Regional Model* could be taken from published tables of the 1970 population and employment census, while the housing data for the Regional Model were provided by the 1968 housing census. The 1968 employment data required for launching the economic submodel of the Regional Model had to be estimated by interpolation using 1961 census data. A major problem in collecting the base year data for the Regional Model was presented by the reorganization of local government in the sixties and seventies which made extensive recalculations necessary in order to transform the data to labor market regions based on current administrative boundaries.

The main data sources for the base year data of the *Zonal Model* were tapes of the 1968 housing census and the 1970 population and employment census specially prepared for this project by the City of Dortmund. They were the basis for establishing the disaggregate (120-type) distributions of households and housing and of the occupancy matrix of the 12 districts of Dortmund. However, for the 19 neighboring communities, for which such tapes were not available, estimates based on one-dimensional distributions taken from statistical tables had to be made. A special estimation technique was developed to substitute the income information not contained in the census data. By this technique each household is associated with one of four income groups depending on the employment status and completed education of its head, both which informations were available on the tapes (cf. Gnad, Vannahme, 1981).

Base year data of the housing stock were taken from the 1968 housing census. As with the household data, tapes containing information on a dwelling-by-dwelling basis were available for Dortmund, while some estimation of distributions had to be made for the neighboring communities. All information needed to establish the 120-type housing distribution for each zone was contained on the tapes. However, the quality attribute had to be estimated as an aggregate of a number of dwelling attributes (cf. Gnad, Vannahme, 1981).

Establishing the base year occupancy matrix presented a special problem. The 1968 housing census contained detailed housing information, but only very limited information about households. The 1970 census contained detailed household, but no housing information. The problem was to match both kinds of census, although they were 18 months apart in time. The problem was solved by first generating for each zone a household-housing matrix from the 1968 data and then «blowing it up» to match the 1970 household distribution (cf. Gnad, Vannahme, 1981).

Base year distribution of industrial and commercial buildings by industrial sector and zone had to be estimated, as presently no such data exist. For this purpose an estimation procedure based on the employment census, the existing land use pattern, and certain assumptions about space requirements of different industries was developed. The existing land use pattern could be retrieved in considerable detail from digitized maps based on aerial photographs.

The data sources for the *District Model* were the same as for the Zonal Model. As a first study area the district of Hombruch (zone 9), one of the largest districts of Dortmund, was selected (cf. fig. 5).

Calibration data for the *Regional Model* are readily available regional employment and population and interregional migration data. Current calibration work focuses on the identification of attributes, weights, and utility functions for the attractiveness indicators in the economic and demographic submodels. Much of this work is experimental in character as the parameters of the attractiveness functions can only partly be estimated statistically. In addition, work is under way to estimate the regional variation of such model parameters as birth rates and labor-participation rates.

The main data sources for calibrating the *Zonal Model* are disaggregate population and housing data of the year 1977 and more aggregate employment, population, housing, and migration data for all other years. The major calibration problems, besides estimation of numerous demographic, technical, and monetary parameters, are connected with the *basic event probabilities* for the aging submodel and the *index of housing satisfaction* for the migration submodel.

The basic event probabilities are partly linked to empirically well established demographic parameters and can be checked against exogenous population projections. Much more difficult is the estimation of probabilities for events like «new household of child», «rise of income», «decrease of income», «retirement», or «new job», for which only few data on the basis of household types exist. However, the only alternative to their approximation by best judgment would be to ignore them, which is no real alternative in a model based so much on household decisions.

Even more crucial is the estimation of the preference functions used to calculate the index of housing satisfaction in the migration submodel. There can be no doubt that the calibration of hundreds of utility functions and weights even for a past period of time, let alone their extrapolation into the future, heavily overtaxes the available data. But again, not to include them in the model would mean to ignore the essential variety of housing needs and tastes, which certainly would be the worse alternative. Consequently, formal estimation techniques in a strict statistical sense play only a minor role in the calibration of the preference functions; instead, many functions are determined by judgment, inferences, analogies, and careful checking of plausibility. The empirical foundations of this informal way of model calibration include the numerous surveys of regional and urban housing markets conducted in the Federal Republic of Germany in recent years.

The main source of calibration data for the *District Model* is the statistics of building completions permanently operated by the City of Dortmund, which contains detailed information about location, time, volume, and building type of all major construction activity in Dortmund. However, as this data source only recently has been automated, information about earlier construction years will be difficult to retrieve. In addition, many minor construction activities, in particular modifications of existing buildings, are usually not recorded at all. Another problem yet to be solved is how to efficiently take into account the multitude of zoning regulations produced during the last decade.

References

Azcarate J. (1970) A Model of a Local Housing Market: SMALA, Proceedings of PTRC Symposium on Housing Models, PTRC, London, 41-49.

Birg H. (1975) Prognose des regionalen Angebots an Arbeitskräften, Sonderheft 105, DIW, Berlin.

Bolton R. (1980) Multiregional Models: Introduction to a Symposium, *Journal of Regional Science, 20*, 131-142.

Gatzweiler H.P. (1975) Zur Selektivität interregionaler Wanderungen, Forschungen zur Raumentwicklung 1, BfLR, Bonn.

Gnad F., Vannahme M. (1981) Haushalts- und Wohnungstypen auf dem Dortmunder Wohnungsmarkt, Working Paper, SFB 26 Münster, Universität Dortmund, Dortmund.

Hall P. (ed.) (1977) *Europe 2000*, Duckworth, London.

Kain J.F., Apgar W.C. Jr., Ginn J.R. (1976) Simulation of Market Effects of Housing Allowances, Vol. I: Description of the NBER Urban Simulation Model, Research Report R77-2, Harvard University, Cambridge, Mass..

Rojahn G. (1981) Prognosen der regionalen Arbeitsplatzentwicklung in Nordrhein-Westfalen, Working Paper, SFB 26 Münster, Universität Dortmund, Dortmund.

Schacht P. (1976) *Ein mikroökonomisches Simulationsmodell zu einem städtischen Wohnungsmarkt - Dargestellt am Beispiel Hamburgs*, Vandenhoeck & Ruprecht, Göttingen.

Schönebeck C. (1981) Ein makroanalytisches Modell der räumlichen Entwicklung von Wirtschaft und Bevölkerung in Nordrhein-Westfalen, Working Paper, SFB 26 Münster, Universität Dortmund, Dortmund.

Schönebeck C., Wegener M. (1977) Kleinräumige Standortwahl und intraregionale Mobilität: die Raum-Zeit-Struktur des Modells, Working Paper, SFB 26 Münster, Universität Dortmund, Dortmund.

Sonnenschein U. (1976) Abgrenzung der Arbeitsmarktregion Dortmund mit Hilfe der Input-Output-Analyse, Working Paper, SFB 26 Münster, Universität Dortmund, Dortmund.

Tillmann H.-G. (1981) Ein mikroanalytisches Modell der räumlichen Entwicklung in einzelnen Stadtbezirken Dortmund, Working Paper, SFB 26 Münster, Universität Dortmund, Dortmund.

van den Berg L., Klaassen L.H. (1978) The Process of Urban Decline, Working Paper 6, Netherlands Economic Institute, Rotterdam.

Wegener M. (1981) Ein mikroanalytisches Modell der Siedlungsentwicklung in der Stadtregion Dortmund, Working Paper, SFB 26 Münster, Universität Dortmund, Dortmund.

Résumé. Cet essai est un compte-rendu d'un projet de recherche mené par l'Institut de Planification Urbaine et Régionale de l'Université de Dortmund, dans lequel les relations entre les changements économiques - sectoriels et technologiques - le choix de localisation, la mobilité et l'utilisation du sol dans les régions urbaines ont été étudiées. A ce but on a construit un modèle dynamique de simulation multi-niveau du développement régional pour simuler les décisions de localisation de l'industrie, des promoteurs résidentiels, des familles, et les effets de ces décisions sur les migrations et les configurations des déplacements. Le modèle simule en outre, les impactes des programmes publiques par rapport au développement industriel, résidentiel et infrastructural.

Le modèle est constitué de trois niveaux spatiaux: (1) un modèle macroanalytique du dévéloppement démographique et économique de 34 régions du marché de la main-d'oeuvre, dans l'état du Nordrhein-Westfalen; (2) un modèle micro-analytique de localisation intra-régionale et des décisions de migration dans 30 zones de la région urbaine de Dortmund, (3) un modèle micro-analytique du dévéloppement de l'utilisation du sol dans un ou plusieurs districts de Dortmund.

L'essai décrit la structure génerale du modèle et de ses principaux niveaux et sous-modèles et présente un résumé des données nécéssaires et des problèmes de calibration rencontrés dans son application pour la région de Dortmund.

Riassunto. Questo articolo è una relazione sul lavoro svolto nell'ambito di un progetto di ricerca condotto presso l'Istituto di Pianificazione Urbana e Regionale dell'Università di Dortmund, nel quale sono state studiate le relazioni tra i cambiamenti economici - settoriali e tecnologici -, la scelta localizzativa, la mobilità e l'uso del suolo nelle regioni urbane. A questo scopo è stato elaborato un modello dinamico di simulazione multilivello dello sviluppo regionale, per simulare le decisioni localizzative dell'industria, dei promotori residenziali, delle famiglie, nonché gli effetti di tali decisioni sulle migrazioni e le configurazioni degli spostamenti. Il modello simula, inoltre, lo sviluppo dell'uso del suolo e gli impatti dei programmi pubblici e delle politiche relativamente allo sviluppo industriale, residenziale ed infrastrutturale.

Il modello è costituito da tre livelli spaziali: (1) un modello macroanalitico dello sviluppo demografico ed economico di 34 regioni di mercato della forza lavoro nello stato del Nordrhein-Westfalen; (2) un modello microanalitico della localizzazione intraregionale e delle decisioni migratorie in 30 zone della regione urbana di Dortmund; (3) un modello microanalitico dello sviluppo dell'uso del suolo in uno o più distretti di Dortmund.

L'articolo contiene una descrizione della struttura generale del modello e dei suoi principali livelli e sottomodelli, nonché un riassunto dei dati necessari e dei problemi di calibrazione incontrati nella sua implementazione per la regione urbana di Dortmund.

Sistemi Urbani, 3, 1982, 403-425 Guida Editori, Napoli, Italia

Models of an urban housing market

B. Hårsman

Stockholm Office of Research and Statistics, Box 2242, S-103 16 Stockholm, Sweden.

Received 30 September 1980, in revised form 13 July 1981

Abstract. The paper contains an overview and describes some applications of a model package developed for analyzing the pattern of housing consumption in the Stockholm region. The framework consists of a household formation model, a housing demand model and of models for depicting the competition between different household categories. The supply and also rents are treated exogenously reflecting the possibilities for the Swedish governments at various levels to control new constructions and rents. Starting from a given population distribution at the regional level the household model uses information about transitions of persons between various ages and household categories to forecast both the total number of households and the numbers belonging to different categories.

In the demand model, housing is viewed as a composite good with different attributes. The market outcome is simulated by a linear programming model and a model based on information theoretical concepts. The former allocates households to dwellings according to differences between bid prices and rents and has been designed to reflect the functioning of a partly controlled housing market of the Swedish type. The latter starts from a priori distribution of the housing demand pattern and uses information about the future number of dwellings and households to derive the most probable a posteriori housing consumption pattern given the information available.

Key words: household, housing, disaggregated demand, application.

1. Introduction

This paper gives an overview of a model package developed for analysing an urban housing market. It also presents some applications with these models (*).

As discussed in earlier papers (Gustafsson *et al.*, 1978; Holm *et al.*, 1976) the models have been designed in close co-operation with the planning authorities of the Stockholm region. This means an orientation towards practical planning problems. It also means that the general setting of the models is a housing market of the Swedish type, i.e. a market where rents are partly controlled and where the volume and structure of new production as well as demolition and modernization are to a great extent controlled by the government and the local authorities.

(*) The models have been developed by an informal research group financed by the Swedish Council for Building Research and the Stockholm municipality. Besides the author, J. R. Gustafsson, F. Snickars, P. Holm and A. E. Andersson are included in the group.

The authority mainly responsible for housing provision in Sweden is the local government. Every year plans for the development of the housing stock during the next five years are prepared. These plans are then used by the national government to decide on upper limits for state loans to housing investment in different regions. Up to 95 per cent of the new dwellings are partly financed by these loans which are granted subject to certain cost and quality standards.

The methods currently used by the local authorities when preparing their housing plans are rather crude. The starting point is usually an estimation of the future demand, obtained through a population forecast and a trend extrapolation of the headship rates. Changes in the existing stock are determined with reference both to observed rates of demolition and renewal and policy goals concerning the quality of housing.

The difference between the estimated number of households and the future housing stock gives the required number of new dwellings. No formal methods are used to determine the composition of new construction. The demand for different types of dwellings and the possibilities of using the system of housing allowances to influence demand are only hinted at and the composition decision is made mainly in the light of current housing consumption and policy goals, i.e. goals concerning dwelling size standards and segregation.

During the seventies the importance of using a microoriented approach when modelling the housing market has been stressed by several authors. According to Kain and Quigley, for example, the American housing market studies performed by the National Bureau of Economic Research show that «the demand for housing and the behaviour of urban housing markets are better understood if «housing» is viewed as bundles of heterogeneous housing attributes rather than as a single-valued commodity, housing services (Kain, Quigley, 1975, p. 1)».

The model package to be presented here is based on such a disaggregate approach. It consists of a household formation model, a model for estimating which dwelling types various household categories will demand and models for depicting the interaction between demand and supply for different dwelling types. The first of the sections to follow outlines the basic model structure and the assumptions made concerning exogenous and endogenous factors. In the next sections the household formation model, the demand model and the market models are successively described. The last section reports on some applications. One type of application shows how the model package can be used to estimate the consumer value of changes in the composition of the housing stock. Another application demonstrates the relation between household formation and housing market conditions, i.g. rents, and the possibilities to use the model package as a population forecasting device.

2. An overview of the models

In fig. 1 an outline is given of the relations between the interlinked models.

Starting with the demand side a crucial question is if the analysis concerns a demarcated housing market or only a part of it. In the

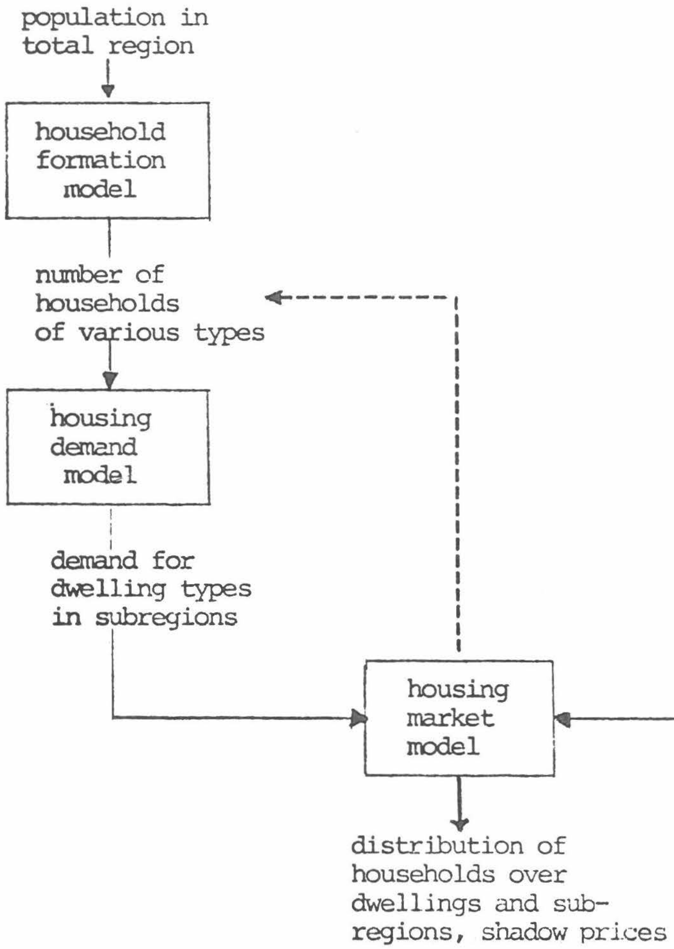


Figure 1 Outline of the relations between the parts of the model package

former case the population development is determined mainly by labour market conditions (provided inmigration is not restricted by an insufficient supply of dwellings) and hence can be viewed as exogenous in relation to changes in the housing market. The population development in an area, which constitutes a part of a larger housing market, is on the other hand to a great extent determined by the development in that market. In both cases then, the analyses must be performed for the total market. The information required about the submarket can of course still be obtained by designing the models so that they include locational aspects.

In this case the population development at the regional level has been taken as given and used as an input to the household formation model. The fact that population changes in a demarcated market can be regarded as independent of the demand for and the supply of dwellings, does not imply that the same holds for the household development. The size and the age distribution of the population are probably basic factors in the household formation process but, as indicated by the dashed line in the figure, the housing market conditions too play an important role. The influence from the market is not explicitly considered in the household formation model. Instead one of the housing market models is used to trace out the relation between household formation and housing market conditions.

The household formation model gives as output the number of households of different sizes and compositions. The main part of this output is used as an input in the demand model which spells out what dwelling types that households in different sizes and income classes demand. In addition to these factors, preferences for various dwelling attributes, rents and housing allowances are also treated explicitely.

In the market model the demand for different dwelling types obtained from the demand model is confronted with a exogenously given supply of dwellings. The rents are also regarded as exogenous variables. Ideally, changes in the housing stock and in rents should be treated endogenously, e.g. as reactions to the market conditions. Here new production, demolition and modernization as well as rents, are instead regarded as instruments the authorities can use to achieve various goals concerning housing comsumption.

3. The household formation model

Between 1945 and 1975 the population of Sweden increased by 23 per cent. Over the same period the number of households, or more precisely the number of households registered as occupants of dwellings, increased by 41 per cent. The population of the municipality of Stockholm has decreased over the latest 15 years but not the number of households. The outlined development is neither unique for Stockholm nor for Sweden.

In view of the social and economic importance of the household structure one would imagine the theoretical framework to be strong. However, that is not the case. Though both demographers and economists have worked with fertility and partner models for many years, a complete «household theory» is still to be developed. As described by Holmberg (1977) the demographic models forming the basis for traditional population forecasts use individuals as building stones and are thus difficult to apply to problems concerning households. Economists like Leibenstein (1957) and Becker (1960) treat family formation as a problem of the optimal number of children and disregard empirical facts like household changes caused by divorces and life cycle traits like the home-leaving of children. The partner models of the kind proposed by e.g. Hoem (1969) are also restricted to one of the aspects of household formation.

The statistical data on household formation are incomplete and scattered. There is considerable information about cross section conditions but only rarely one may obtain data concerning time series. Many data, at least in Sweden, are adapted for use in forecasting the total number of households by the classical headship ratio method. If time series exist they do not reveal gross flows of individuals between household categories. Family statistics about marriages and divorces are unreliable since households may often be formed by non-married couples. Fertility and mortality data are tied to individuals although family and household formation is, to say the least, crucial for the first of these processes.

A basic reason for the facts listed above is that household formation is regarded as being a difficult topic in itself. There is a large number of conceivable household categories and the dynamics of the household formation processes are complicated. Households are like nuclear particles. They may split up and rejoin into a wide variety of combinations.

For the above-mentioned reason some basic decisions have been taken concerning the current household formation model. It should be usable to forecast both the total number of households and the number of households of different sizes and compositions. It should be simple and not necessarily rest on micro-oriented analyses. It should also, at least implicitly, be general enough to capture the significance of marriages, divorces and the other aspects of household formation discussed above.

One possible approach to fullfill these requirements is to regard the household formation as flows of individuals between different household categories. As outlined in the following table the current, or future, number of households of a certain category can be seen as determined by the inflow and outflow of persons to that category during the preceding time period (*).

(*) It should be noted that the household formation model presented here has led to the collection of this kind of data.

The row sums show the number of persons belonging to different size classes in 1970 and the number of persons born and outmigrated during the period 1970-75. For example 479 000 persons belonged to households consisting of one or two persons in 1970 and 190 000 persons moved into the Stockholm region during the period 1970-75. Correspondingly the column sums show the distribution of persons among size classes in 1975 and the number of deaths and outmigrants during the period 1971-75.

1975 / 1970	1 and 2 persons	3 and 4 persons	5+ persons	Dead 1971-75	Outmigrants 1971-75	Total
1 and 2 persons	335	51	3	43	47	479
3 and 4 persons	151	346	32	10	68	606
5+ persons	31	67	97	2	30	227
Born 1971-75	5	63	15	1	13	97
Inmigrants 1971-75	55	49	17	9	61	190
Total	557	576	164	64	219	

Table 1 Transitions of persons, in thousands, between households of different sizes in the Stockholm region 1970-75

The rest of the table shows the gross flows behind the development of the household structure. The rows give the destination of the flows and the columns the origin of the numbers observed in 1975. For example, 151 000 of the 606 000 persons belonging to households consisting of 3 or 4 persons in 1970 moved to a smaller size class, 346 000 remained in the same category and 32 000 moved to a larger size class. Of the 576 000 persons who belonged to households consisting of 3 or 4 persons in 1975 the main part, 346 000, were in the same size class in 1970. 112 000 persons were «new» - born or inmigrants - 51 000 came from smaller and 67 000 from bigger households.

A structuring of the household formation process along the lines demonstrated in the table can be used in several ways. By classifying the individuals according to age, for example, one can elucidate the home-leaving by children and by a classification according to sex the formation of couples can be analysed. As shown by Hårsman, Snickars (1979) the flows of persons between household categories can also be included as an additional feature in population models of the transitional type.

A straightforward way of using transitions of the kind illustrated in table 1 is to compute the corresponding transition probabilities. If these probabilities are multiplied with the latest observed number of persons in different household categories then both the population and the household development can be forecasted. In a traditional population forecast, trends in fertility and mortality rates as well as labour market conditions, are normally handled explicitly. The influence of these factors on household formation may be captured by adjusting the calculated transition probabilities to a given population forecast. This can be done in the following way.

Let Q_{ij}^{kl} denote the number of persons belonging to household category i and age group j at the beginning, and to household category k and age group l at the end of the observation period. If P_{ij}^{kl} denotes the corresponding flows during the forecasting period, and N^l the future number of persons in age group l according to the population forecast, the following efficient information adding procedure may be used (*).

$$\underset{P_{ij}^{Kl}}{\text{Minimize}} \quad \sum_i \sum_j \sum_k \sum_l P_{ij}^{kl} \ln(P_{ij}^{kl} / Q_{ij}^{kl}) \tag{1}$$

$$\text{subject to} \quad \sum_k \sum_l P_{ij}^{kl} = \sum_m \sum_n Q_{mn}^{ij} \tag{2}$$

$$\sum_i \sum_j \sum_k P_{ij}^{kl} = N^l \ . \tag{3}$$

According to (1) the information added when transforming the observed flows Q_{ij}^{kl} to the future flows P_{ij}^{kl} should be minimized. As shown by (2) the flows P_{ij}^{kl} must be consistent with the number of persons in different categories in the beginning of the forecasting period, i.e. the end of the observation period. By (3) they are also adapted to the population forecast.

Hobson (1969) has shown that this kind of transformation is an efficient way of using available information and Jaynes (1957) has characterized the forecast P_{ij}^{kl} values as «maximally non-committal with regard to missing information». As shown by Snickars, Weibull (1977), the solution obtained is also in a certain sense the most probable one.

When applied in the Stockholm region the procedure given by (1)-(3) was used in a step-wise fashion to forecast the household development 15 years ahead. The households were classified into five size classes and, within each one of these four categories showing the composition regarding children and youth. Eight age groups were used for persons belonging to these household categories and none for persons in the groups «born», «dead», «inmigrants» and «outmigrants».

(*) For simplicity the groups «born», «dead», «inmigrants» and «outmigrants» are here regarded as household categories.

One way of improving the described forecasting method is of course to extend the classification scheme. Another is to insert more information into the restriction system. For example knowledge about when children leave home or at which ages people usually get married could be used. However, since the flows in the model comprise five years this is not an easy task. A third approach would be to use exogenous information about the relation between household formation and economic factors. As will be shown in a following section such information may be produced by a housing market model. If (1)-(3) is complemented with restrictions of the following type

$$H_i - \Delta H_i \leq \sum_k \sum_l w_{kl} \left(\sum_i \sum_j P_{ij}^{kl} \right) \leq H_i + \Delta H_i \tag{3'}$$

this information can be used. Here $H_i \pm \Delta H_i$ denotes an interval for the future number of households in category i obtained from the housing market model, and w_{kl} are weights reflecting the size of the various household categories.

4. The demand model

In an earlier paper (see Gustafsson *et al.*, 1977), it has been demonstrated how Alonso's bid price concept and Luce's choice theory can be employed to construct housing demand functions. According to Luce (1959) the probability of choosing a consumption bundle c from a given set C of alternatives can be expressed as

$$p(c) = v(c) / \sum_{c' \in C} v(c') \tag{4}$$

where v(c) indicates the utility of consumption bundle c. The elements of the vector c designate different housing attributes and the level of other consumption. Equation (4) tells us that the probability of choosing a certain dwelling depends on the relative valuation of attributes and level of other consumption. This rather general observation may be sharpened by specifying an utility function. To do this the following notation is used,

z_{ik} = dummy variable indicating level k, $k = 1 \ldots m_i$, of housing attribute i, $i = 1 \ldots m$.

y = income

r = rent

a_{ik}, ρ = preference parameters.

Equation (5) gives a rather general utility function which is also feasible for empirical testing.

$$v(z,y-r) = e^{\sum_i \sum_k a_{ik} z_{ik}} (y-r)^{\rho} \ . \qquad (5)$$

If the housing market is supposed to be in equilibrium, consumption data can be used to estimate the preference parameters, see for example Wheaton (1972). Because of rent control and a relatively high level of market inertia this condition is, however, not fulfilled in Sweden. One remaining possibility is to seek direct information about the preference for different housing attributes via a household survey. This is the approach used here. In a survey in the Stockholm region 1975 (see Hårsman, 1976), households were confronted with a set of hypothetical choices between the current dwelling and dwellings that differed with respect to one or several characteristics. By definition, indifference occurred when the chance was 50 per cent that the household would prefer the alternative to the reference dwelling. As illustrated in table 2 the corresponding rent level, i.e. the bid price for the alternative was found by iterating over choice probabilities.

Rent increase (Sw kr)	Probability of choosing the alternative dwelling (per cent) 0	25	50	75	100
0					X
500	X				
100					X
400		X			
200				X	
300		X			

X denotes that the household has given this answer

Table 2 Example of bid price response sequence for a question concerning one extra room, ceteris paribus, not regarding moving costs

The questions were posed to the household in the order shown in the table where also one possible outcome is given. In the example the household obviously has its indifference point somewhere between rent increases of 200 and 300 sw.kr.

The set of questions concerning different attribute changes posed to a household was determined with reference to the characteristics of the current dwelling. The questions concerned both single and composite attribute changes. Moving costs were estimated separately as local attachment and they are thus not reflected in the bids for other attributes. The identification of the attachment level was accomplished by offering a dwelling similar to the current one in all respects but located in another part of the region. The following attributes and attribute levels were used in the survey.

z_1 = dwelling size, measured as number of room units, i.e. kitchen included;

z_2 = standard, indicating whether the dwelling is modern or obsolete according to the technical fittings. A dwelling is defined as «obsolete» if one or more of the following equipment components are missing: bathroom or shower, central heating, water toilet, hot water, sewage. Otherwise it is defined as «modern»;

z_3 = house type, indicating whether the dwelling is an owner occupied house, or in a block of flats. Since dwellings in multi-family houses usually are rented this variable at the same time distinguishes between the corresponding tenure-ownership forms;

z_4 = local density, measuring the degree of green space in the vicinity of the dwelling (within walking distance). A three step scale is used: namely «compact city», «medium density» and «garden city»;

z_5 = location, measured as total return travel time from the dwelling to the CBD an average day;

z_6 = indicator of whether a dwelling is in the current residential district or not (local attachment);

z_7 = other consumption (disposable household income minus rent).

According to (4) the probability $p(c', c)$ that a household prefers an offered consumption bundle c' to a current bundle c can be written as

$$p(c',c)/[1-p(c',c)] = v(c')/v(c) \ . \tag{6}$$

Using the utility function (5) and rewriting in logarithmic form we have

$$\ln \frac{p(c',c)}{1-p(c',c)} = \sum_i \sum_k a_{ik}(z'_{ik} - z_{ik}) + \rho \ln \frac{y-r'}{y-r} \tag{7}$$

where $p(c', c)$ indicates the probability of choosing an alternative dwelling c' instead of the current dwelling c.

This is the so called logit functional form we have used in the empirical work. It should be noted that this function is additive, introducing independency among attributes. However, it is possible to generalize it to include interaction terms.

The estimations have been performed for subgroups of households with similar characteristics as regards household size and income using a maximum likelihood technique. This technique is usually applied when estimating logit models describing patterns of actual choices, e.g. Domencich, McFadden (1975). In such cases the outcomes of choices are used to estimate the underlying choice probabilities. Here, however, we are concerned with actual observations of the (subjective) choice probabilities and the approach to take is not entirely obvious. When logit models are used to analyse consumption data, the error term in the function estimated is usually assumed to be generated by random preference differences within the subgroup of individuals or households being studied. In this case one may instead assume that each of the subjective probability values obtained in the interviews suffers from an error term reflecting the hypothetical character of the choice situation. If the «true» subjective probabilities are identical for households, in a given socio-economic group facing identical choice situations, there is no need, in the estimations, to distinguish this case from the one where actual choices have been observed.

Equation (7) relates the choice probabilities to attributes and preferences in a binary choice situation. Applying (4) the obtained parameter estimates can be used to compute multiple choice probabilities. Since $p(c', c) = 0{,}5$ corresponds to $v(c') = v(c)$ equation (7) can also be used to compute bid prices, i.e. the maximum amounts various households are willing to pay for different dwelling types, while remaining on the same utility level.

Combining these bid prices and choice probabilities with the above described household formation model gives the demand side of the housing market models presented in the following section.

5. Housing market models

Two models have been developed to depict the competition in the housing market. One is a linear programming model which allocates households to dwellings according to differences between bid prices and rents. The other starts from an a priori distribution of the housing demand pattern and uses information about the future number of dwellings and households to derive the most probable a posteriori housing consumption pattern given the information available, i.e. it is the same approach as the one used in the household formation model.

5.1. *The linear programming model*

The model employed is of the Herbert, Stevens (1960) type but the design differs somewhat from the free market picture given in e.g. Anas (1973). Here the aim will primarily be to study the impact of various policy measures and of changes in factors uncontrolled by a regional housing authority.

The following notations are needed.

W = consumer surplus

X_{ij} = number of households in category i allocated to dwellings of type j

p_{ij} = bid price for households in category i for dwellings of type j

b_{ij} = housing allowance for households in category i living in dwellings of type j

r_j = rent for dwellings of type j

H_i = number of households in category i

D_j = number of dwellings of type j

X^o_{ij} = number of households in category i currently living in dwellings of type j

$\gamma_{ij}X^o_{ij}$ = number of households in category i living in dwellings of type j that do not move from their present dwelling.

If the households are assumed to maximize their utility the interaction between housing demand and housing supply may be depicted with the following model:

$$\underset{X_{ij}}{\text{Maximize}} \quad W = \sum_i \sum_j (p_{ij} + b_{ij} - r_j)X_{ij} \tag{8}$$

subject to

$$\sum_j X_{ij} = H_i \tag{9}$$

$$\sum_i X_{ij} \leq D_j \tag{10}$$

$$X_{ij} \geq \gamma_{ij}X^o_{ij} \; . \tag{11}$$

The maximization of total consumers' surplus, equation (8), should be performed subject to the conditions that all households must be accomodated, constraint (9), and that dwelling capacity is not exceeded, constraint (10) (*). If the H_i-variables are regarded as the potential

(*) Thus the total number of households should not exceed the total number of dwellings.

number of households an inequality sign can be used in (9). This will be further commented in a following section. Equation (11) is inserted to reflect the inertia in the housing market. It has been designed to allow various assumptions to be tested for different household categories and dwelling types.

Using the model one can analyze how the optimal consumption pattern depends on bid prices - and hence for example on incomes-housing allowances, and rents as well as on the number of different households and dwellings. The effects of marginal changes in housing supply or number of households are given directly through the dual. Disregarding equation (11) the dual problem can be formulated as follows.

$$\underset{u_j, v_i}{\text{Minimize}} \quad \sum_j u_j D_j + \sum_i v_i H_i \tag{12}$$

subject to

$$u_j + v_i \geq r_{ij} + b_{ij} - r_j \tag{13}$$

$$u_j \geq 0 \tag{14}$$

$$v_i \gtrless 0 \,. \tag{15}$$

Here u_j and v_i are the dual variables associated with the number of dwellings of type j and households of category i, respectively. The variable u_j indicates how much the maximum consumer surplus given by (8) is changed when a dwelling of type j is added to the housing stock, and v_i indicates the corresponding change when a household of category i is added.

The u-variables may be interpreted as scarcity prices and the v-variables as a subvention system. If the rents r_j are set according to production and running costs the dual can be given an interesting interpretation. As shown by Hårsman, Snickars (1975) the goal function (12) can then be seen as the profit of a planning body fixing rents to $r_j + u_j$ and using a subsidy system $b_{ij} - v_i$. When this profit is minimized the consumer surplus given by (8) is maximized.

5.2. *The efficient information adding model*

Let X'_{ij} denote the forecasted number of households of category i demanding dwellings of type j. Using (7) the following expression can be derived for X'_{ij}

$$X'_{ij} = H_i \frac{a_{ij}(y_i - r_j)^{\rho_i}}{\sum_k a_{ik}(y_i - r_k)^{\rho_i}} \tag{16}$$

where H_i denotes the forecasted number of households of category i and a_{ij} summarizes the composite attribute value of dwelling type j for households of category i. Equation (16) necessitates the use of data of the kind described in section four. An alternative way of estimating X'_{ij} is given below.

$$X'_{ij} = H_i x^o_{ij} \ . \tag{17}$$

Here x^o_{ij} represents the share of households in category i currently living in dwellings of type j, i.e. $x^o_{ij} = X^o_{ij} / \sum_i X^o_{ij}$.

The problem of forecasting the future consumption pattern X_{ij} may be posed as a question of how to transform the demand distribution X'_{ij} and at the same time use the information given by equation (9) and (10′)

$$\sum_j X_{ij} = H_i \tag{9}$$

$$\sum_i X_{ij} = D_j \ . \tag{10'}$$

As in the LP-model, constraint (9) means that all households of category i must have a dwelling. The reason for using equality sign in the capacity constraint (10) is purely technical (*).

As shown in section three the transformation problem can be solved by minimizing the information added, i.e.

$$\underset{X_{ij}}{\text{Minimize}} \sum_i \sum_j X_{ij} \ln (X_{ij} / X'_{ij}) \tag{18}$$

subject to constraints (9) and (10').

Using Lagrange techniques the solution may be expressed analytically as given by (19)

$$X_{ij} = u'_i v'_j X'_{ij} \tag{19}$$

where u'_i and v'_j are monotone transformations of Lagrange multipliers related to constraints (9) and (10′), respectively. Hence the forecasted number of households belonging to category i living in dwelling of type j equals a corrected value of the corresponding demand. The correction factor, $u'_i v'_j$, reflects the competition among the households.

It should finally be pointed out that entropy maximizing is in a sense a special case of this model. When the distribution X'_{ij} is uniform, reflecting no a priori information, (18) becomes the classical entropy measure.

(*) In applications vacant dwellings have been included in X'_{ij} and X_{ij}.

6. Applications of the LP-model

The purpose of this section is to illustrate the possibilities of using the described model package. This is done by means of impact-oriented applications with the LP-model. Two types of applications will be presented. The first concerns analysis of how the consumer surplus in the Stockholm region housing market depends on the supply of single family houses. A detailed description of this study - which was performed at request of the Stockholm Regional Planning Office - is given by Hårsman, Snickars (1980). In this application the model given by (8)-(11) is used with the exception of the allowance system, i.e. the b_{ij} variables. In the other type of application the relation between household formation and housing market conditions is studied and a system of housing allowances is introduced. The constraint that all households must have a dwelling, equation (9), is here replaced by (9′)

$$\sum_j X_{ij} \leq H_i \ . \qquad (9')$$

All model applications concern the Stockholm region. This region is for the actual purpose divided into two major parts: the municipality of Stockholm and the surrounding municipalities. Five classes are used for dwelling size, two for type of structure and three for internal equipment quality. This gives a total of 60 dwelling types. Table 3 gives a summary of the supply of these dwelling types in 1975 (*).

	Stockholm	Surrounding municipalities	Total region
Single family houses of which			
$\geq$ 4 ru	10 800	29 300	40 100
$>$ 4 ru	26 200	86 800	113 100
Multi family houses of which			
$\leq$ 4 ru	263 000	168 200	431 200
$>$ 4 ru	34 000	30 900	64 900
Total	334 100	315 200	649 300

Table 3 Number of dwellings according to type of structure and number of room units (ru), Stockholm, surrounding municipalities and Stockholm region in 1975

(*) The table only includes dwellings occupied by «registered» households, i.e. 92 per cent of the total supply.

The supply of dwellings is somewhat larger in Stockholm than in the surrounding municipalities. We can also see that small dwellings in multi-family houses dominate the supply in Stockholm. In the remaining part of the Stockholm region the dwellings are larger and single family houses constitute approximately one third of the total dwelling supply. As shown in table 4 this supply structure is reflected in the household composition (*):

	Stockholm	Surrounding municipalities	Total region
1 person	153 400	82 700	236 100
2 persons	104 800	88 900	193 700
3 persons	40 200	58 400	98 700
4 persons	26 000	59 500	85 500
5+ persons	9 600	25 700	35 300
Total	334 100	315 200	649 300

Table 4 Number of households of different sizes, Stockholm, surrounding municipalities and Stockholm region in 1975

The small households dominate both in Stockholm and in the other municipalities. Table 4 also shows that small households are concentrated to Stockholm and large households in the rest of the region.

The households within each size class have been classified according to disposable household income. For each of the 34 household categories obtained in that way bid prices have been computed for the 60 dwelling types. This was done in three steps. Estimations of equation (7), - see Gustafsson *et al.* (1978), - were used to compute how much a household in a certain category, *living in a certain dwelling type*, bids for various alternatives. In a second step these price bids were checked against income and if necessary corrected to guarantee a subsistence level. Using the current consumption pattern the resulting bids were finally weighted together to obtain 60 price bids for each one of the 34 household categories.

(*) A household is as mentioned above defined as a so called dwelling household.

6.1. *Consumer surplus - percentage number of single family houses*

The model applications have been constructed around a basic alternative, corresponding to the current number of households and dwellings. In order to make the alternatives «realistic» they reflect the possibilities to change the percentage number of single family houses during a five-year period. The alternatives are given in table 5.

	Basic alternative (1975)	Alternative		
		I	II	III
Single family houses	153	122	178	208
Multi family houses	496	527	471	441
Total supply	649	649	649	649
Percentage number in single family houses	23,6	18,6	27,4	32,1

Table 5 Thousands of dwellings in single family and multifamily houses in the Stockholm region 1975, according to three alternatives

The proportion of single family houses, which currently is 23,6 percent, ranges between 18,6 and 32,1 percent in the alternatives.

It is not self evident which values one should choose for the γ_{ij}-variables. There is little knowledge about mobility propensities for households. However, approximately 50 percent of all households in the Stockholm region have lived 5 years or more in their current dwellings. As a rough approximation of the mobility during five years a value of 0,5 has been used for all γ_{ij}-variables. The basic alternative has also been run with all values equal to zero. Furthermore, the effects of an increased number of small households and of a 20 percent rent increase for single family houses have been analysed (*).

Table 6 shows how the consumer surplus depends on the proportion of single family houses.

	Proportion single family houses			
	18,6%	23,6%	27,4%	32,1%
Consumer surplus	9	35	41	46

Table 6 Consumer surplus in million sw. kr per month in the Stockholm region 1975

(*) The rent for a single family house is an estimated rent equivalent.

The consumer surplus increases at a decreasing rate when the proportion of single family houses is increased. In the basic alternative (23,6%) the surplus equals 50 sw. kr per household and month, which constitutes seven per cent of the average rent per month. The scarcity prices change according to what one could expect when the proportion of single family houses increases: they decrease for single family houses and increase for dwellings in multi-family houses.

In table 7 the effects are shown of 20 per cent higher rents for single family houses and of an increased proportion of small households. The result of using γ_{ij}-values equal to zero is also given.

	Consumer surplus
Basic alternative	35
Higher rents for single family houses	34
Increased proportion of small households	31
Maximal mobility alternative ($\gamma_{ij} = 0$)	108

Table 7 Consumer surplus in million sw. kr per month in the Stockholm region 1975

As expected the surplus is somewhat lower in a housing market with higher prices for single family houses or more small households. An increased number of small households produces higher scarcity prices for small dwellings and lower scarcity prices for large dwellings. The alternative with more expensive single family houses is associated with a strong decrease in the corresponding scarcity prices, which explains the marginal change in the surplus. A market without inertia produces a very high surplus. A comparison between such a market and the basic alternative indicates very high moving costs: approximately 110 sw. kr per household and month. This high level indicates the importance of considering moving costs explicitly in housing market studies. A policy conclusion is that efforts to increase mobility in the housing market probably have a high pay-off.

6.2. *Household formation - housing market conditions*

As already mentioned the household formation is studied by defining H_i as a potential upper limit for the number of households of category i. Unless otherwise stated the γ_{ij}-variables equal 0,5. A discussion of the impact on household formation of rent variations, housing allowances and changes in the supply of dwellings is performed in this section.

Table 8 shows the number of households «produced» at the current rent level and when rents increase by 10 and 20 per cent, respectively. As a comparison the current number of households is included.

	Stockholm	Surrounding municipalities	Stockholm region
Current number	334	315	649
Model computations			
Current rents	296	288	584
10% rent increase	285	293	578
20% rent increase	285	279	564

Table 8 Thousands of households in Stockholm, surrounding municipalities and the Stockholm region. Current numbers and model alternatives

The computed numbers are considerably lower than the current values. One reason for these deviations is that the model is not calibrated to be a forecasting tool. Another reason is that housing allowances and other means to support low income households are not included. Similarly, a *rough* picture is given of the relation between household formation and the rent level. When rents are increased the number of households in the region decrease. The elasticity seems to range from $-0{,}1$ to $-0{,}2$. The household decrease is slower in the surrounding municipalities than in Stockholm. In fact, the number of households in the surrounding municipalities increases when the rent increase is 10 per cent. The corresponding population redistribution is still stronger. Hence the results to some extent support the hypothesis that increased housing costs tend to concentrate the population in an urban region.

As shown by table 9 the impact of a rent increase differs considerably between small and large households.

	Current number	Model computations		
		Current rents	+ 10%	+ 20%
1 person	236	210	199	185
2 persons	194	183	180	180
3 persons	99	97	95	95
4 persons	86	65	75	77
5+ persons	35	30	30	27
Total	649	584	578	564

Table 9 Thousands of households by size in the Stockholm region. Current numbers and model alternatives

The 1-person household seems to be most sensitive to a rent increase. It is tempting to conclude that this reflects difficulties for young people to move from their parents when the rent level is high.

Though rather complex, the current Swedish system of housing allowances is in principle constructed in the following way,

$$b = f(y, n, r) \tag{20}$$

where b denotes the subsidy, y income, n number of persons in the household and r rent. For households qualifying for the subsidy, b increases when n or r increase and decreases when y increases (*). Two sets of b_{ij}-values, B_1 and B_2, are used to simulate this subsidy system. In B_1 low-income households obtain higher subsidies, and high-income households smaller subsidies as compared to B_2. Because of poor statistical data it is impossible to say which alternative resembles the current system best. However, the total cost of the housing allowance system in the Stockholm region in 1975 was approximately 55-65 million sw.kr per month, which constitutes 12-15 per cent of the rent payments. Depending on whether B_1 or B_2 are applied, the corresponding cost is 76 or 70 million sw.kr per month, respectively. Hence both B_1 and B_2 are more generous than the current system. In table 10 the number of households obtained by choosing B_1 and B_2 values are compared with the current numbers and an alternative without subsidies (i.e., the «current rents» alternative in tables 8 and 9).

	Stockholm	Surrounding municipalities	Stockholm region
Current number	334	315	649
Model computations			
No subsidies	296	288	584
B_1	294	296	590
B_2	290	299	589

Table 10 Thousands of households in Stockholm, surrounding municipalities and the Stockholm region. Current numbers and model alternatives

The cost of the subsidies is 80 million sw.kr per month in alternative B_1 and 76 million sw.kr per month in alternative B_2. This constitutes 21 and 19 per cent of the corresponding rent payments, respectively. At the same time as the model market uses the system of housing allowances more effectively than the real one, it produces more

(*) The subsidy is only given to households with incomes below and rents above certain limits.

households. In spite of the difference in design, the impact on household formation is similar in both alternatives. It is also interesting to note that the effects in Stockholm are reverse to those in surrounding municipalities. This indicates that a system of housing allowances has a dispersing effect, i.e. the opposite effect of a rent increase.

In a market where rents are partly controlled, the supply of dwellings plays an important part in household formation. Both the total supply and the supply structure can be assumed to have an influence. For example, an increased supply of small dwellings ought to be positively correlated with the formation of small households. In order to illustrate these effects an alternative supply has been tested. This alternative corresponds to changes in the housing stock during a three-year period, i.e. to the supply of dwellings in the Stockholm region as of 1972. In table 11 the results of using this supply and current rents are compared to the «current supply and current rents» alternative.

No. of room units	Increase in no. of dwellings Abs.	%	No. of persons in the household	Increase in no. of households Abs.	%
– 2	5 500	4	1	4 300	2
3	8 200	5	2	12 700	7
4	11 400	7	3	2 600	3
5	7 300	9	4	8 500	15
6 +	10 800	14	5 +	500	2
Total	43 200	7	Total	28 400	5

Table 11 Household and supply changes by size in the Stockholm region

The 43 000 new dwellings result in 28 000 new households. The corresponding percentage figures are 7 and 5, respectively, which means that the household elasticity with respect to supply is 0,7. Although the additional supply of new dwellings is dominated by large units, the size of the majority of new households is small.

The rent increase analysed earlier in this section also indicated larger changes among small as compared to large households. Hence, the «pure» market conditions seem to be most significant for the formation of small households. As also demonstrated above the introduction of a housing allowance system of the Swedish type has a reverse effect, to the market effect. The net impact of these opposite tendencies has not been analyzed here. Even though partial, it seems however safe to conclude, that model applications of the kind presented can be used to enhance our understanding of the household formation process.

References

Anas A. (1973) A Dynamic Disequilibrium Model of Residential Location, *Environment and Planning, 5*, 633-647.

Becker G. (1960) An Economic Analysis of Fertility, in *Demographic and Economic Changes in Developed Countries*, Princeton University Press, New York.

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

Gustafsson J.R., Hårsman B., Snickars F. (1977) Housing Models and Consumer Preferences: Applications for the Stockholm Region, *Papers of the Regional Science Association, 38*, 125-147.

Gustafsson J.R., Hårsman B., Snickars F. (1978) Medium Term Forecasting Models for the Housing Market: the Stockholm Case, *Papers of the Regional Science Association, 40*, 87-106.

Hårsman B. (1976) Housing Preferences and Housing Demand-Results from an Interview Survey in Stockholm 1975, The Master Planning Commission, Stockholm.

Hårsman B., Snickars F. (1975) Disaggregated Housing Demand Models: Some Theoretical Approaches, *Papers of the Regional Science Association, 34*, 121-143.

Hårsman B., Snickars F. (1979) A Method for Disaggregate Household Forecasts, Paper presented at a Regional Science Association Meeting, Los Angeles.

Hårsman B., Snickars F. (1980) The Demand for Single Family Housing, Stockholm, (mimeo in Swedish).

Herbert J., Stevens B. (1960) A Model for the Distribution of Residential Activity in Urban Areas, *Journal of Regional Science, 2*, 21-36.

Hobson A. (1969) A New Theorem of Information Theory, *Journal of Statistical Physics, 1.*

Hoem J. (1969) A Theoretical Discussion of the Effect of Selection of Life Style on Mortality for Married and Unmarried Men, Central Bureau of Statistics, Oslo (in Norwegian).

Holm P., Snickars F., Gustafsson J.R., Hårsman B. (1976) Policy-Oriented Housing Models - Some Tentative Applications, *London Papers in Regional Science*, Pion, London, 150-170.

Holmberg I. (1977) Demographical Models of Population Change: Outline of Development and an Appraisal, in Andersson Å.E., Holmberg I. (eds.) *Demographic, Economic and Social Interaction*, Ballinger, 25-53.

Jaynes E. (1957) Information Theory and Statistical Mechanics I, *Physical Review, 106*, 171-190.

Kain J., Quigley J. (1975) *Housing Markets and Racial Discrimination: a Microeconomic Analysis*, Columbia University Press, New York.

Leibenstein H. (1957) *Economic Backwardness and Economic Growth*, Chapman and Hall, London.

Luce R. (1959) *Individual Choice Behavior*, Wiley, New York.

Snickars F., Weibull J. (1977) A Minimum Information Principle-Theory and Practice, *Regional Science and Urban Economics, 7*, 137-168.

Wheaton W. (1972) Income and Urban Location, Ph. D. Dissertation, University of Pennsylvania, Philadelphia.

Résumé. Cet essai présente une revue et la déscription de quelques applications d'un package de modèles dévéloppés pour analyser la structure des consommations de logements dans la région de Stokholm. Le cadre de référence est constitué d'un modèle de formation des familles, d'un modèle de demande de logements et de modèles qui décrivent la compétition parmis différentes catégories familiales. L'offre et les rentes résidentielles sont traitées d'une façon exogène à fin de tenir compte de la possibilité du Gouvernement Suédois de contrôler, aux différents niveaux, les nouvelles constructions et les rentes. Se basant sur une distribution donnée de la population au niveau régional, le modèle de la formation des familles utilise les informations relatives aux transitions de personnes entre les différents âges et les catégories familiales pour prévoir soit le nombre totale de familles soit le nombre de familles appartenant aux différentes catégories. Dans le modèle de la demande, l'habitation est considéré comme un bien composé ayant différents attributs. La solution du marché est simulé par un modèle à programmation linéaire et par un modèle basé sur des conceptions de la théorie de l'information. Le premier de ces modèles assigne les familles aux logements se basant sur les différences entre les prix de l'offre et les rentes, et a été conçu de façon à tenir compte du fonctionnement d'un marché de logements, partiellement contrôlé de les authorités publiques, comme celui en Suède. Le deuxième modèle se base sur une distribution à priori de la structure de la demande de logements et utilise les informations sur le nombre futur de logements et de familles pour obtenir la plus probable structure à posteriori des consommations de logements étant données les informations disponibles.

Riassunto. Questo articolo presenta una rassegna e la descrizione di alcune applicazioni di un package di modelli sviluppati per analizzare la struttura dei consumi di abitazioni nella regione di Stoccolma. Il quadro di riferimento è costituito da un modello di formazione delle famiglie, da un modello di domanda di abitazioni e da modelli di descrizione della competizione tra differenti categorie familiari. L'offerta residenziale e le rendite residenziali sono trattate esogenamente al fine di tener conto delle possibilità del Governo Svedese di controllare, ai vari livelli, le nuove costruzioni e le rendite. Partendo da una data distribuzione della popolazione al livello regionale, il modello di formazione delle famiglie utilizza le informazioni relative alle transizioni di persone tra le diverse età e le categorie familiari per predire sia il numero totale delle famiglie sia il numero di famiglie appartenenti alle diverse categorie. Nel modello di domanda, l'abitazione è vista come un bene composito con differenti attributi. La soluzione di mercato è simulata da un modello a programmazione lineare e da un modello basato sui concetti della teoria dell'informazione. Il primo di detti modelli assegna le famiglie alle abitazioni, sulla base delle differenze tra i prezzi di offerta e le rendite, ed è stato concepito in modo da riflettere il funzionamento di un mercato delle abitazioni, in parte controllato dall'operatore pubblico, quale è quello svedese. Il secondo modello parte da una distribuzione a priori della configurazione della domanda di abitazioni ed utilizza le informazioni sul numero futuro delle abitazioni e delle famiglie per derivare la più probabile configurazione a posteriori del consumo di abitazioni, date le informazioni disponibili.

Sistemi Urbani, 3, 1982, 427-452 Guida Editori, Napoli, Italia

Environmental quality, abatement and urban development

U. Schubert
Interdisciplinary Institute for Urban and Regional Studies, University of Economics, Hasenauerstrasse 42, Vienna, Austria.

Received 30 September 1980, in revised form 15 April 1981

Abstract. The purpose of this paper is to derive a simultaneous equation model of an urban economy in which the impacts of environmental policy on the suburbanization-reurbanization process can be assessed. The variable to show these effects is a measure of the density of land-use activity. The model consists of essentially four components - an urban land market, a market for an aggregate good, a description of a process of emission of residuals as a necessary by-product of land-use activities (production, consumption, commuting), and their accumulation and diffusion in a receiving medium (e.g., air and noise pollution).
Environmental policy is then introduced into this simultaneous equation model (emission standards and fines, ordinances to install abatement equipment, emission fees, etc.). A discussion follows on the hypothesis that a successful antipollution policy could introduce a trend toward reurbanization with the claim that clear air could induce land users to relocate closer to the center of a city.
The policies considered include fines for exceeding emission standards, inducing land users to purchase abatement devices in the market. The study attempts to derive the impacts of these measures on residuals' concentration, the roles of abatement devices, and urban density based on some *a priori* assumptions.

Key words: environmental quality, emissions, density, abatement.

1. Introduction

Several land value studies (such as Anderson, Crocker, 1971) undertaken in highly industrialized countries show that environmental quality plays an ever increasing role in the choice of location by urban land users, particularly households and some (usually service-oriented) firms. The most important factors of environmental quality in an urban area are usually the levels of air and noise pollution, which are the effects of the levels and density of land-use activities.

Low concentration of residuals - a stock variable - can be found mostly in the suburban areas of a city. The flow variables determining the stock of residuals at a given location are emissions (depending on activity levels and density), and the diffusion and absorption by the environment. As residual density is usually lower in the suburbs, we expect its concentration to decrease as the distance from the city center increases.

The concentration of residuals, among other factors, has led to an accelerated suburbanization process, the results of which are ever increasing energy costs, inner city problems, and rising total pollution (van den Berg, *et al.*, 1981).

Suppose now that a city authority takes steps to reduce emissions in order to improve environmental quality within the city's boundaries. (In this paper the measures applied are not differentiated by location.) If this policy turns out to be successful, *one* of the incentives to locate in the suburbs - better environmental quality - has vanished. As the location closer to town become more accessible and transportation costs become smaller, there may be a side effect of environmental policy leading to a process of reurbanization. In a continuous process of income growth, however, where larger lots and new homes play the essential role in the location decision, it is to be suspected that the claimed effect will be over-compensated again.

In this paper an analytical model, based on *a priori* reasoning, is proposed that assesses the effects of various policy measures on population density. Since the model is basically a static one, only comparative static results are obtained (Schubert, 1979). We will restrict ourselves to the assessment of the impact of an urban environmental policy that attempts to curb the emissions of pollutants in the urban area by means of obbligatory or voluntary installation of abatement devices. It should be added at this point, that the model outlined is, in principle, empirically testable. A thorough econometric analysis has so far been prevented by the lack of data, especially on policy measures. (Some preliminary first tests based on Austrian data have been carried out, but will not be reported here).

2. A model of environmental quality in an urban area

2.1. *Basic assumptions*

Six basic assumptions are adhered to in our model.

- Following the von Thünen-Alonso (von Thünen, 1826; Alonso, 1964) analysis, we assume the model city to be situated on a homogeneous plain.
- We then postulate a «neutral» residual-receiving medium (i.e. in the case of air and noise pollution, no predominant direction of winds).
- All transactions take place in the center of town where consumers and producers have to go for their «transacting». (Strictly speaking, since all firms are not in the center of town, we assume that the «labor market» is in the central business district [CBD], and that workers commute first to this area and then to their place of work. Although this assumption may not always be realistic, it does simplify the analysis considerably).
- To determine the activity levels of producers and consumers, we assume that households maximize utility, given an income constraint, and producers maximize profits, given a production function.

- The decisions made by the city government remain exogenous throughout the analysis.
- Our final assumption is that the land market is an example of monopolistic competition. This implies that locations are differentiated by their characteristics and that bargaining establishes their price (following Alonso's «game theoretic» approach).

2.2. *Components of the model*

Concentration of residuals (R)

Environmental quality can be described by the distribution of residuals [R(r)] over the urban area (in the stationary state). The quality will depend on *total emissions*, which are caused by the *land-use activities'* consumption, production, and commuting to the center, and on the physical processes' «diffusion» and «absorption».

The urban land market (q)

The supply of land in a given zone is fixed (in a circular city: $2r\pi$). Demand for land is derived from utility and profit maximization. Among other variables it will depend on environmental quality and accessibility of the CBD.

The goods market (X)

To facilitate the analysis, the total supply of consumption goods X, is assumed to be produced in the urban region (activity levels of firms = supply of X). This supply function (derived from the profit maximum conditions) depends again on accessibility and for some firms on environmental quality. It also depends on the factor market conditions, which are exogenously given. Demand depends on income (exogenous) and, indirectly, on environmental quality and accessibility (among other variables).

Emissions

All land-use activities cause emissions, which are considered as the part of waste that is «harmful» (i.e., which causes negative externalities), to be deposited in a common property resource where they accumulate. For our analysis we consider the predominant property resource to be air.

The city's administration

Urban planning in this context plays two roles: that of providing the transportation infrastructure (thus influencing accessibility) and that of applying various antipollution schemes. The activities of the administration are not space-consuming in the context of this paper, thus urban land users are only households and firms.

2.3. *The micro-economic background of the demand for land and the determination of activity levels*

Potential land users evaluate offers on the real estate market by assessing the different characteristics of the various offers. Out of the entire range of such possible characteristics, we will only consider two «broad categories» (Richardson, 1978): accessibility and environmental quality. «Location» is seen as a good produced essentially by two agents, the city administration and all land users (Bökemann, 1977).

Accessibility is «produced» by the urban planner, but when certain capacity thresholds are surpassed and congestion begins to be a problem, all commuters contribute to «accessibility». It will be measured, therefore, in terms of travel time from a land user's location to the city center (A).

Environmental quality is «produced» by the emissions of all urban land users, these being caused by the land-use activities. The city administration plays an indirect role by applying environmental policy measures. Environmental quality, will be measured by some aggregate index of various noxious residuals (R).

Households

The utility of a household is given by:

$$u = u(X, q, A, R)$$

its income constraint:

$$y = p^X X + p^q q + TC \qquad E = E(X, r)$$

where

y : income

X : level of aggregate consumption good

q : size of lot

NP : pollution abatement devices

E : emissions (caused by consumption of X and commuting a distance of r to the center)

p^X, p^q: prices

TC : transportation costs

r : distance from the center.

Equilibrium of the household is found by maximizing u, subject to y, where A and R are exogenously given, varying with r. Setting the first order conditions equal to zero and solving for X and q yields demand functions for X and q, depending on p^X, p^q, y, TC, A and R (we postulate that the second order conditions hold):

i.e. $q^D = q^D(p^q, p^X, y, TC, A, R)$

and $X^D = X^D(p^X, p^q, y, TC, A, R)$.

We expect the following «sensitivities» to changes in these variables (they cannot, as usual, be rigorously derived from the common micro-economic regularity conditions):

$$q^D_{p^q},\ q^D_{TC},\ q^D_{p^X} < 0$$

$$q^D_y,\ q^D_A,\ q^D_R > 0$$

and similarly

$$X^D_{p^X},\ X^D_{p^q},\ X^D_{TC} < 0\ , \qquad X^D_y,\ X^D_A,\ X^D_R > 0\ .$$

Firms

Profits: $P = p^X X - (wL + p^q q + TC)$.
The production function is:

$$X = f(L, q, A, R)$$

$$E = E(X, r)$$

where:

X: output of consumption goods
L: non-land production factors
p: prices
w: wage rate.

The first order conditions of a profit maximum can be solved to yield a supply function for X and demand functions for q and L. (As the factor market will be assumed to be in equilibrium, it will not be analyzed any further).

$$X^S = X(p^X, p^q, w, TC, A, R)$$

$$q^D = q^D(p^q, p^X, w, TC, A, R)$$

where:

$$X^S_{p^q},\ X^S_w,\ X^S_{TC},\ X^S_A > 0$$

$$X^S_R \leq 0$$

$$X^S_{p^X} > 0\ .$$

2.4. *Emissions and the distribution of residuals over the urban area*

There are three land-use activities causing emissions:

- Production and consumption of the «aggregate» good (activity level: X).
- Commuting to the center of the city.
- Transactions at the center (which we will assume to be constant, so they will be left out of the analysis).

Activities constitute transformation processes - «inputs» are turned into «outputs» - some of which can be used further (*«consumption goods»*) and *waste.* In terms of physical mass, the total mass of inputs is equal to the mass of outputs. Production activities use production factors (raw materials, energy, etc.) which are transformed into goods and services and waste.

Households «consume» goods - i.e., they are being transformed into waste. Commuting takes energy - which is also turned into waste. Some of the waste, however, is recycled or is harmless, some of it is noxious, some of it could be treated and transformed into something directly useful or better «digestible» for nature.

The *noxious* part of *waste* we will define as «emissions» (pollution, residuals).

$E^X = E(X)$ emissions caused by activities production or consumption of X

$$E^{commuting} = E^{co} = E(r) \text{ (or, with congestion),}$$
$$E^{co} = E(A;\ A \text{ is traveling time}).$$

Once residuals are emitted into a common property resource they *accumulate* and *diffuse.* Some of them are transformed by nature into harmless materials (regeneration), thus reducing the stock of pollutants in the receiving medium. Hence, we have the following stock-flow relationship:

$$\frac{dR}{dt} = \dot{R} = E^T - (c)^2R,\ (c)^2 > 0\ ,\ \text{constant}$$

where:

R : stock of residuals

E^T : total emissions of residuals (flow)

$(c)^2$: rate of natural regeneration (c is squared for mathematical convenience)

t : time.

As this total stock of residuals spreads over space, we observe varying environmental quality levels at different locations. Looking at the movements of pollutant particles after emission we observe that they move randomly in all directions from the source of emission. As we have excluded any permanently disturbing factors in the medium, a certain amount, say b, of the residuals always moves to the «left», the same amount to the «right» of the source (we are talking about a one-dimensional space at the moment): $-bR_r$ moves in the «positive» and bR_r in the «negative» direction (Feller, 1980).

At a given point in time, then, the change in the stock of residuals at a given location can come from other locations (diffusion), or from emissions (at the location), or be due to regeneration (at the location). The total change in residuals within an interval of dr then becomes

$$\dot{R}dr = b\,[R_r(r+dr;t) - R_r(r,t)] + E^T(r,t) - (c)^2 R(r+dr;t)\,dr \ .$$

To obtain the steady state distribution of residuals over space we let $t \rightarrow \infty$, i.e. $\dot{R} = 0$ (time disappears as a variable):

$$b\,\frac{d^2R}{dr^2} + E^T(r) - (c)^2R(r) = 0 \ .$$

Leaving the one-dimensional space and turning to our circularly symmetric city, we can define location in terms of polar coordinates - but as the angle does not matter (symmetry!) we finally obtain (setting $b = 1$, without loss of generality)

$$\frac{d^2R(r)}{dr^2} + \frac{1}{r}\,\frac{dR(r)}{dr} - c^2R(r) = -E^T(r)$$

(for details, see Schubert, 1979). We assume $R(0)$ and $\lim_{r \rightarrow \infty} R(r)$ to be finite, and we are considering total emissions $E^T(r)$ to be E^X and E^{co}.

Now let there be d land users in a given ring (with distance from the center r). As each land user in this ring has activity level X, and emissions E(X), to obtain total emissions in the ring we have to sum over all the land users, or to simplify, we compute: E(X) d = total emissions due to X in ring r (where d(r) is the number of land users in a given ring r, and E(X) are the average emissions/land user). The emissions rate density then is

$$\frac{E[X(r)\,d(r)]}{2r\pi}$$

as the total area in a given ring is $2r\pi$.

Each land user on his way to the center emits $E^{co}(r)$ in each ring. How many are passing through a given ring r? Everybody located outside r, towards the edge of town contributes to total emissions at r.

There are, say, N(r) land users passing thorugh r (where N is the sum of all land users d outside r).

Total commuting emissions: $E^{co}(r)\,N(r)$

Emissions rate density: $\frac{E^{co}(r)}{2r\pi}N(r)$.

Note that E(X) and $E^{co}(r)$ are (provisionally) exogenous in this partial analysis. The solution of the differential equation yielding the partial distribution of residuals over space has to be of the form:

$$R = R'[E(X),\ E^{co}(r),\ d,\ N(d),\ r]$$

or $R = R(X, d, r)$.

It can be shown that the solution is of the form:

$$R = [E^{co}\int_0^r a\,N(a)\,I_0(ca)\,da]\,K_0(cr)$$

$$E(X)\,I_0(cr)\int_r^{r_m} a\,N(a)\,K_0(ca)\,da$$

where r_m is the «edge» of the urban region and a is an auxiliary variable and I_0 and K_0 are two standard, tabulated Bessel functions, which can be roughly drawn as in fig. 1 (Abramovitz, Stegun, 1972).

(Assuming emission-rate densities declining with distance from the center, the slope of R will resemble a Gaussian).
It can be shown that $R_X, R_d > 0$ and $R_r < 0$, given a decreasing emission-rate density toward the periphery of the city. As this density (given our assumptions) varies negatively with a rising r, the above assertions will hold. (For details, see Schubert, 1979).

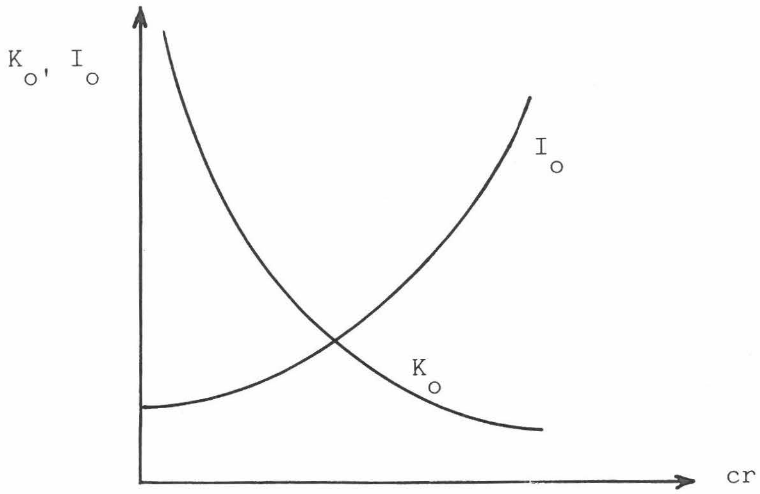


Figure 1 Components of the steady state distribution of residuals

3. Market equilibrium

3.1. *Density and the land market*

The total supply of land in a given ring is $2r\pi$. Demand is equal to $q^D = q^D(y, w, p^q, p^X, A, R, TC)$, for all urban land users. Bargaining will eliminate all those potential land users that do not offer a high enough price at a given location. The equilibrium price of land will hence be the simultaneous solution of supply and demand equations - i.e. land prices p^q will be a function of all exogenous (to the land market) variables.

$$p^q = p'(A, R, y, w, p^X, TC, 2r\pi)$$

but as $A = A(r)$, $TC = TC(r)$, and y and w are exogenous, we can write: $p^q = p(R, r, z)$,

where: $z = (y, w, \pi, p^X)$.

The equilibrium solution of the land market will also inform us about the number of land users located in a given ring

i.e.: $d = \frac{2r\pi}{q^D}$.

3.2. *The goods market*

Individual demand for and supply of X depends on p^X as well as the other exogenous variables. Market equilibrium implies that the price of X, the solution of the simultaneous demand = supply system, depends on the exogenous variables defining supply and demand. Thus $p^X = p(R, r, z)$. The equilibrium quantities of X equal the activity levels causing emissions. (To make this step, we had to neglect all imports into and exports out of the urban region considered. The city's economy constitutes a closed system). The equilibrium activity levels of urban land users hence become

$$X = X(y, w, p^q, p^X, A, R, TC)$$

where

$$X_y > 0, X_w, X_p^q, X_{TC} < 0 \ .$$

X_A and X_R cannot be assessed *a priori* - households demand more X when environmental quality (accessibility) drops, firms are either not affected at all (by R) or have a decreased output.

4. A simultaneous model of markets and residuals concentration in a city

We have been discussing a feedback system, in which some variables were taken as provisionally exogenous to derive partial equilibrium conditions and the stationary solution of the economic-environmental space/time process (where time is considered «short-run» in this formulation) (see Isard, Liossatos, 1978).

Residuals concentration depends on activity levels of land users and their density - determined in the land and goods market, which in turn depend on environmental quality.

$$R = R(X, d, r) \qquad (1)$$

$$d = \frac{2r\pi}{q^D(p^q, p^X, R, z, r)} \qquad (2)$$

$$p^q = p^q(R, z, r) \tag{3}$$

$$X = X(p^X, p^q, R, z, r) \tag{4}$$

$$p^X = p^X(R, z, r) \ . \tag{5}$$

As this model constitutes a simple general equilibrium type of model (with two goods involved), we can use Walras's law, i.e., relative prices. Thus

$$p^q/p^X = p$$

where we can set $p^X = 1$, without loss of generality. We can now substitute p for p^q and p^X in all equations and drop (5).

Is an equilibrium solution for this system of implicit, simultaneous equations defined; i.e., are the endogenous variables R, d, p and X defined in terms of the exogenous variables z and r? Using the «implicit function theorem», we have to postulate *continuity* of all implicit functions and the existence of *continuous derivatives*. These conditions are fulfilled by assumption. Furthermore, the Jacobian determinant $|J| \neq 0$.

Let us rewrite the equations in implicit form:

$$F_1: R(X, d, r) - R = 0$$

$$F_2: \frac{2r\pi}{q(R, p, z, r)} - d = 0$$

$$F_3: p(R, z, r) - p = 0$$

$$F_4: X(p, R, z, r) - X = 0.$$

Table 1 shows [J], the derivatives of these functions with the signs in parantheses.
The term in brackets is unambiguously positive, the second negative (assuming that X_R is negligible in quantity). The strength of the positive effect depends on r, as the magnitude of the first expression varies with r. There is, hence, in general a positive and a negative branch of the Jacobian and a point where it vanishes, i.e., where changes in the exogenous variables do not affect the endogenous variables. To the left and the right of this point, effects have opposite signs (see also figs. 2 and 3).

	R		d		p		X	
F^1	-1		R_d	$(+)$	0		R_X	$(+)$
F^2	$\frac{2r\pi}{q^2} q_R$	$(-)$	-1		$\frac{2r\pi}{q^2} q_p$	$(+)$	0	
F^3	P_R	$(-)$	0		-1		0	
F^4	X_R	$(?)$	0		X_p	$(-)$	-1	

$$|J| = \left[1 + R_d \frac{2r\pi}{q^2} (p_R q_p + q_R)\right] - R_X(p_R X_p + X_R)$$

$$\begin{matrix} + & (-)(-) & & + (-)(-) \sim o \\ & + \quad + & & \end{matrix}$$

Table 1 The derivatives of the simultaneous, implicit form equations

5. The impact of environmental policy measures on the «geography» of a city

5.1. *Assessing density changes in the model*

In the previous section a simultaneous model of urban land and goods markets was introduced. «Environmental quality», the result of the land-use decisions and physical processes, was introduced into the evaluation and decision calculus of land users.

Comparing the results of this feedback system to Alonso's (1964) and Muth's (1969), we observe that it is most likely that a positive valuation of environmental quality tends to «stretch» the city - a process of suburbanization. The price and density gradients (the equilibrium solutions) tend to flatten pushing the «edge» of the city outward (see Schubert, 1979).

Suppose now that the city administration attempts to reduce pollution in the urban area. Since the deterioration of environmental quality in locations close to the city center was one of the driving forces of suburbanization, it could well be that a reversal of this process could be the result of environmental policy (van den Berg *et al.*, 1981; Edel, 1972). To substantiate this claim, we must isolate certain effects, i.e., distinguish between the income growth effect behind suburbanization (leading to demand for bigger lots) and the environmental factor, which drives people out of town and into areas of lower residuals concentration - regardless of the size of their individual lots.

How can the effects of environmental policy on urban shape be demonstrated? We will use the «density of land-use gradient», to show the spatial effects of environmental policy. To facilitate the exposition of the claimed hypotheses we make some simplifying assumptions about the shape of d. Let d be the usual bell-shaped function as shown in fig. 2. The integral of this function is the total population of the urban area. (As our city is symmetric we can use a two dimensional curve for illustration.)

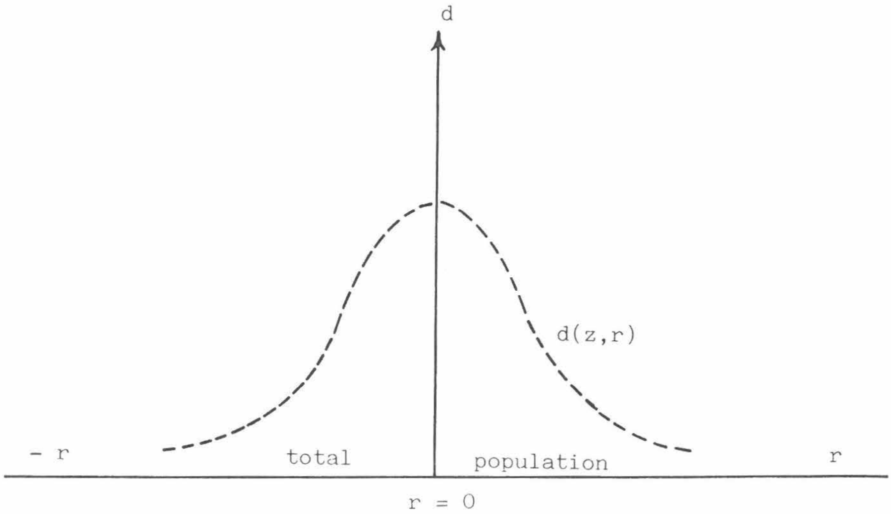


Figure 2 The distribution of population over urban space

Comparing density profiles we can call a city «more compact», if the total population lives in a smaller area. We assume that the *total* population of our model city does not change. As some policy variables are altered, only the distribution over space may change.

A compact city has more people living around the center than does a «dispersed» city (see fig. 3). In equilibrium d depends on r and the other exogenous variables (among them the instruments of environmental policy). Let this set of variables be represented by the vector z. We have then

$$d = d(r, z) .$$

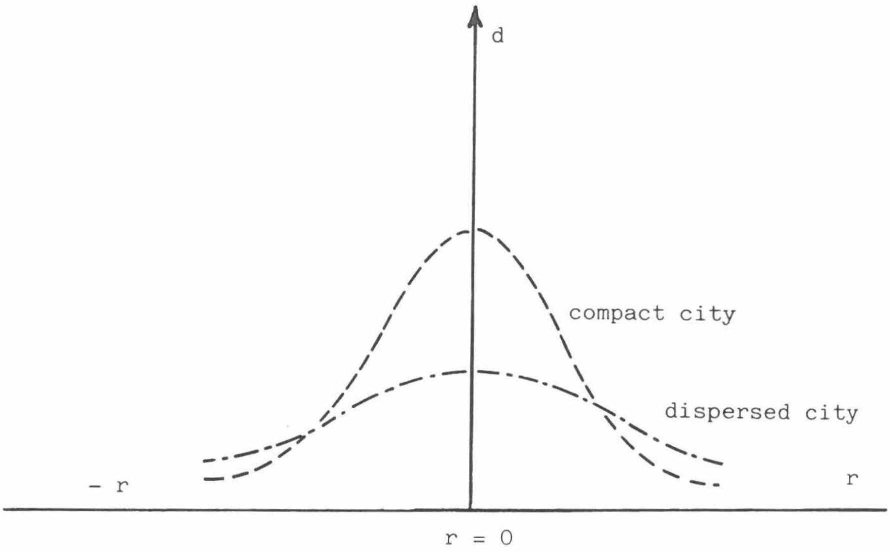


Figure 3 Density profiles of different urban forms

Changes in z will cause shifts in d. We claim now that, given our assumptions about d, a compact city is characterized by a «smaller» standard deviation of the population distribution, as is illustrated in fig. 4.

Our task is to show how σ changes with z. To facilitate the derivation, «characteristic functions» will be used. Instead of using the standard deviation σ we can use the second moment m_2 of the

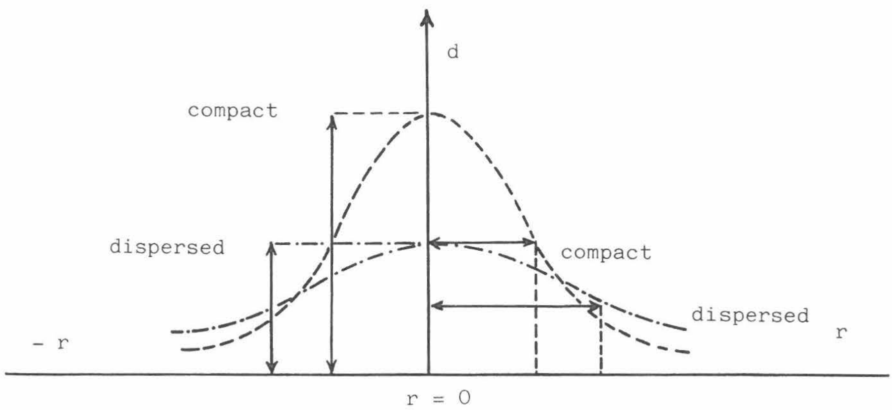


Figure 4 Urban form in terms of the standard deviation of density profiles

distribution d to show the impact of changes in the policy variables z without loss of information (Fisz, 1973). To substantiate the claim of a reurbanization effect of environmental policy, we have to show that

$$\frac{\partial m_2}{\partial z} < 0 .$$

A characteristic function Φ of a density function is defined as (Fisz, 1973)

$$\Phi(\lambda, r, z) = \int_{-\infty}^{\infty} d(r, z)\, e^{i\lambda r}\, dr$$

where λ is an auxiliary variable, and $i = \sqrt{-1}$.

The second moment (m_2) of this functions is

$$m_2 = -\frac{\partial^2 \Phi}{\partial \lambda_2} \quad (\text{at } \lambda = 0)$$

so we obtain

$$m_2 = + \int_{-\infty}^{\infty} d(r, z)\; r^2 dr . \tag{6}$$

Differentiating m_2 with respect to z we get

$$\frac{\partial m_2}{\partial z} = \int_{-\infty}^{\infty} d_z r^2 dr \left(d_z = \frac{\partial d}{\partial z} \right) . \tag{7}$$

We claim that

$$\int_{-\infty}^{\infty} d_z r^2\, dr < 0 .$$

Since we postulate that the population remains unchanged in size and only changes its distribution over space, we must also have

$$\int_{-\infty}^{\infty} d_z dr = 0 .$$

The integral over the urban area constitutes the population. The population difference when policy variables are changed must be equal to zero. But if the area under the curve remains the same, the density functions with or without z have to intersect (see also figs. 3 and 4). Before the intersection point, toward $r = 0$, the city is more compact if its density curve lies above the curve of the dispersed city. After the intersection point, toward the periphery, the reverse must hold.

For the compact city, hence, we must observe that toward the center, policy must shift d upwards, i.e., $d_z > 0$ at $r \to 0$. At the intersection point, policy measures do not affect the density curve at all, i.e., $d_z = 0$; outside the intersection points then $d_z < 0$ must hold. (For details see Schubert, 1979). What has to be demonstrated in the following then is

$$d_z \gtrless 0$$

as r changes. We will do this by trying to show that $d_z > 0$ at $r \sim 0$. The general shape of d_z will also give some indication of the possible adjustment processes set off by environment policy in all the markets and in the residuals concentration relation.

5.2. *Environmental policy and abatement*

In this paper we are dealing with abatement policies by the urban authorities. In principle, there are two kinds of abatement. *Emissions* can be controlled by purchasing and operating equipment (e.g., exhaust fume filters in automobiles, noise insulation on lawn mowers, etc.), or the effects of *residuals concentration* can be mitigated (e.g., insulation in houses against outside noise, private air purification, etc.). But not all policies provide an incentive *to invest in antipollution devices.*

Without explicit policy (e.g., Ruff, 1972; Schubert, 1973) individual abatement against emissions is highly unlikely. Individuals will purchase equipment to the extent that the marginal benefits from a lower concentration of residuals exceeds the marginal cost of buying and operating it. A control of emissions is hence only likely in the case of individuals whose own emissions constitute a good part of the ambient residuals concentration experienced by them. (If one's *own* noise and smoke is the most severe environmental problem, one will do something about it - which is hardly ever the case at higher densities of land use, where it is mostly the neighborhood effects that determine environmental quality. This is even more true of commuting pollution.)

We will not explicitly analyze changes in production and consumption technology (abatement sets in *after* the waste has been produced). In order to do so, a disaggregated model would be necessary differentiating between different kinds of technologies, inputs, goods, and transportation modes. Emissions in our model are seen to depend only on the total activity level.

In order to assess the impact of some environmental policies within our simultaneous system, the model has to be extended. First, a market for «abatement factors» (NP) has to be added; specifically we will have to analyze the impact of environmental policy on the demand for NP. The price p^{NP} will be treated as exogenous. Most of these measures have consequences on other markets as well, usually via the «income effect».

5.3. *A market for abatement devices*

Consumers ad producers can purchase devices to reduce emissions. (Again note that waste cannot be reduced due to the law of conservation of mass, but harmful residuals, defined as «emissions» in this analysis, can be reduced.) In equilibrium the total demand for abatement devices has to be equal to their supply.

Total demand: $NP^D = NP^D_{\text{firms}} + NP^D_{\text{households}}$

Total supply: NP^s, exogenous.

Hence, we have $NP^D = NP^S = NP$. The price of these devices, p^{NP}, is considered exogenous, NP^D can be found by looking at the individual decisions of land users. In partial equilibrium, NP as one of the (endogenous) decision variables is a function of the exogenous variables.

In this case this implies that NP = Function (exogenous variables, parameters). We will analyze the impacts of various policies on this partial equilibrium solution first and then assess the «system-impact». There remains a problem, however. Firms supplying abatement devices also pollute. (Their activity levels are NP.) When all production takes place in the urban area the negative effect of this production could outweigh the positive effect via filtering, etc.. Looking at R now, it becomes

$$R = R(X, NP, d, r).$$

But what is the effect of NP on R? The abatement effect decreases R, but the production of NP increases it. To facilitate the analysis, we will assume that $\partial R / \partial NP = R_{NP}$ has a negative net effect on residuals concentration. ($R_{NP} < 0$). (The diffusion process demonstrated in Section 2 will not be explicitly changed by introducing NP, the implications of which are easily seen.)

Our simultaneous model now becomes

$$p^q/p^X = p^1, p^{NP}/p^X = p^2$$

$$F^1: R(X, NP, d, r) - R = 0$$

$$F^2: \frac{2r\pi}{q(p, z)} - d = 0$$

$$F^3: p^1(R, r, z) - p^1 = 0$$

$$F^4: X(p^1, R, z) - X = 0$$

$$F^5: NP(p^2, R, z) - NP = 0.$$

In terms of endogenous variables as used before, the z stands for a vector of exogenous variables to be analyzed in the following sections; p is a vector of relative prices.

$$R_X, R_d > 0, \ R_{NP} < 0$$

$$q_p < 0, \ q_R > 0$$

$$p_R < 0$$

$$X_p < 0, \ X_R = 0 .$$

The micro economic background

Maximum allowable emission rates are set ($\bar{E}$). Violations of the set standards are fined. This fine rises linearly with the emission excess, i.e., total fine: $(E-\bar{E})F$, where $F > 0$, and $E^X = E(X, NP)$ or $E^{co} = E(r, NP)$, or F is an emission fee to be paid per unit of emitted residuals. The household's decision problem is

$$\text{Max.} \quad u(X, q, A, R)$$

$$\text{s.t.} \quad \text{(i)} \ y - (p^X X + p^q q + TC + F(E-\bar{E}) + p^{NP} NP)$$

$$\text{(ii)} \ y - (p^X X + p^q q + TC + FE + p^{NP} NP) .$$

Assuming second-order conditions to hold, we can find the first derivatives of the appropriate Lagrangian and set them equal to zero, which will yield a system of simultaneous equations.

We can solve these simultaneous equations implicitly by defining X^E, q^D, and NP^D; these solutions will be defined in terms of prices (p^X, p^q, p^{NP}), and the exogenous variables TC, a, $\bar{E}$, F, $2r\pi$. How do changes in the exogenous variables affect the demand for X, q, and NP? As some of the effects have been outlined before, we will only look at the following:

$$NP_F \quad \text{and} \quad NP_{\bar{E}}\,,\ NP_{pNP}\,,\ NP_A\,,\ NP_R\,,\ NP_{TC}$$

$p_F \geq 0$ (as the fine increases it pays to abate more)

$$NP_{\bar{E}} \leq 0\,,\ NP_{pNP} < 0\,,\ NP_A = 0\,,\ NP_R \sim 0\,,\ NP_{TC} < 0\,.$$

As long as $E < \bar{E}$, there will be no effect of F or $\bar{E}$. The analogous problem for the urban firms is

$$\text{Max.} \quad P = p^X X - (p^q q + wL + p^{NP} NP + F(E-\bar{E}) + TC)$$

$$\text{s.t.} \quad X - f(L, q, A, R) = 0\,.$$

For the resulting equations we can again determine the supply of X and the demand for L, q and NP in terms of prices and the exogenous variables. We need now

$$NP_{pNP} < 0 \qquad NP_R \sim 0$$

$$NP_A = 0 \qquad NP_{TC} < 0$$

$$NP_F = 0 \qquad NP_{\bar{E}} < 0\,.$$

Total demand equals demand by households (NP_H^D) plus firms (NP_F^D). Let $NP^S = NP^D = NP$; then NP is the (partial) equilibrium solution of the demand=supply condition in the market. It is defined as a function of the exogenous variables, i.e.,

$$NP_H^D(p^X, p^q, p^{NP}, A, TC, \bar{E}, F, y) +$$

$$+ NP_F^D(p^X, p^{NP}, p^q, w, A, R, TC, F, \bar{E}) = NP^S\,.$$

From this we get

$$NP = NP(p^{NP}, p^X, p^q, w, A, y, R, TC, F, \bar{E})\,.$$

Effects of abatement demand in the simultaneous model

Adding the implicit equation to the simultaneous model implies the following structure (using $p^q/p^X = p$ again, setting $p^X = 1$):

$$F^1:\ R(X, NP\,d, r) - R = 0$$

$$F^2:\ \frac{2r\pi}{q(y, w, p, p^{NP}, R, A, TC, F, \bar{E})} - d = 0$$

$$F^3:\ p(y, w, p^{NP}, R, A, TC, F, \bar{E}) - p = 0$$

$$F^4:\ X(y, w, p^{NP}, p, R, A, TC, F, \bar{E}) - X = 0$$

$$F^5:\ NP(y, w, p^{NP}, p, R, A, TC, F, \bar{E}) - NP = 0$$

from which the following Jacobian matrix can be derived (the signs of partial derivatives are in parentheses). As in table 1, we note that the Jacobian consists of a positive and a negative branch and a point where |J| vanishes. The location of this «turning point» of the signs of effects depends on r, of which J will be positive for small values, and negative for large ones.

R		d		p		X		NP	
-1		R_d	(+)	0		R_X	(+)	R_{NP}	(−)
$\frac{-2r\pi}{q^2} q_R$	(−)	-1		$\frac{-2r\pi}{q^2} q_p$	(+)	0		0	
P_R	(−)	0		-1		0		0	
X_R	(?)	0		X_p	(−)	-1		0	
NP_R	(0)	0		NP_p	(−)	0		-1	

Table 2 The derivatives of the simultaneous, implicit form equations with a market for abatement equipment added to the model

$$|J| = -1 - R_d \frac{2r\pi}{q^2} (p_R q_p + q_R) + R_X (X_R + p_R R_{NP} NP_p X_p)$$

$$\underset{(<0)}{-\ +}\qquad +\quad (-)(-)(+)\qquad \underset{(>0)}{+\sim 0\ \ (-)(-)\quad (-)(-)}\ .$$

Let us first see whether there will be any demand for pollution abatement equipment in equilibrium when fees (fines) are charged on emissions; i.e., $NP_F^+ > 0$

$$NP_F^+ = \frac{|J_{NP(F)}|}{|J|} \quad \text{by Cramer's rule}$$

$$\begin{aligned}
|J_{NP(F)}| = & - F^2_F R_d (NP_R + NP_p p_R) \\
& \quad (+)(+)\ (+) \quad (-)(-) \\
& + F_F^3 \{ R_X (X_p NP_R + NP_p X_R) \\
& \quad (-)\ (+)(-)(+) \quad (-)\ (\sim 0) \\
& - (NP_p - R_d \frac{2r\pi}{q^2} q_p NP_R + NP_p \frac{2r\pi}{q^2} q_R R_d) \} \\
& \quad (-) \quad (+) \quad (-)(+) \quad (-)\ (+) \quad (+) \\
& - F_F^4 R_X \quad (NP_R + NP_p p_R) \\
& \quad (-)(-)(+)\ (+)\ (-)\ (-) \\
& + F_F^5 R_X (X_R + X_p p_R) \\
& \quad (+)(+)(\sim 0)\ (-)(-) \\
& - (1 + R_d \frac{2r\pi}{q^2} q_p p_R + \frac{2r\pi}{q^2} q_R R_d) \\
& \quad (+) \quad (-)(-) \quad (+)\ (+)
\end{aligned}$$

as $F_F^1 = R_F = 0$

$$F_F^2 = -\frac{2r\pi}{q^2} q_F \quad (+)$$

$$F_F^3 = P_F \quad (-)$$

$$F_F^4 = X_F \quad (-)$$

$$F_F^5 = NP_F \quad (+) .$$

Let us look at the components of this expression individually.

F^2 - Collecting emission fees (fines) tends to increase density directly, as less land can be afforded by land users. The higher density deteriorates environmental quality, which makes land prices fall - but this makes it possible to purchase more antipollution devices.

F^3 - As income has to be spent on emission fees, less can be paid for land, but this makes demand for NP go down, thus worsening environmental quality. This could affect activity levels and consequently R. Lower land prices permit the purchase of more abatement equipment; R decreases. This leaves land users satisfied with smaller lots, which makes density go up, but at the same time increases residuals.

F^4 - Expenditures for X decrease with rising emission fees; consequently there is less emission and environmental quality rises. Land prices now go up, which leaves less income to be spent on abatement.

F^5 - As emission fees become higher, land users attempt to emit less to save and the demand for NP shifts upwards. More abatement implies better environmental quality, a possible change in activity levels, and thus the concentration of residuals. As R goes down land prices rise, decreasing activity levels, and further improving environmental quality.

The higher land prices also increase density, causing a deterioration of environmental quality. Density tends to increase also as R decreases because of more abatement and land users demand less land, thus improving environmental quality further. The chains of effects in F^2, F^3, and F^5 support the hypothesis that the introduction of emission fees (fines) tend to encourage the installation of antipollution devices; F^4 points in the opposite direction. Letting $r \rightarrow 0$ leaves us with $|J_{NF(F)}| > 0$.

As $|J|$ and $|J_{NP(F)}|$ are positive at locations close to the center, while the reverse holds for suburban locations, the introduction of fines (fees) will cause a positive demand for abatement devices. (There is a small zone where there is no demand at all.) Does this induced demand for antipollution devices tend to decrease the second moment of the residual concentration function? We will again use d_F^+ for our argument.

$$d_F^+ = \frac{|d(F)|}{|J|} .$$

We replace the second column of $|J|$ by

$$
-\begin{bmatrix} F_F^1 \\ F_F^2 \\ F_F^3 \\ F_F^4 \\ F_F^5 \end{bmatrix} = -\begin{bmatrix} 0 & \\ -\dfrac{2r\pi}{q^2} q_F & (+) \\ p_F & (-) \\ X_F & (-) \\ NP_F & (+) \end{bmatrix}
$$

$$
|J_{d(F)}| = (-1) \Big[\underset{(+)}{F_F^2} (\underset{(-)}{R_{NP}} \underset{(-)}{NP_p} \underset{(-)}{p_R} + \underset{(+)}{R_X} \underset{(\sim 0)}{X_R} + \underset{(+)}{R_X} \underset{(-)}{X_p} \underset{(-)}{p_R})
$$

$$
+ \underset{(-)}{F_F^3} \underset{(+)}{\frac{2r\pi}{q^2}} (\underset{(+)}{q_R} \underset{(-)}{R_{NP}} \underset{(-)}{NP_p} + \underset{(+)}{q_R} \underset{(+)}{R_X} \underset{(-)}{X_p} - \underset{(+)}{R_X} \underset{(\sim 0)}{X_R} \underset{(-)}{q_p} - \underset{(-)}{q_p})
$$

$$
+ \underset{(-)}{F_F^4} \underset{(+)}{\frac{2r\pi}{q^2}} (\underset{(-)}{q_R} \underset{(-)}{p_R} \underset{(+)}{R_X} + \underset{(+)}{q_R} \underset{(+)}{R_X})
$$

$$
+ \underset{(+)}{F_F^5} \underset{(+)}{\frac{2r\pi}{q^2}} (- \underset{(-)}{p_R} \underset{(-)}{R_{NP}} + \underset{(+)}{q_R} \underset{(-)}{R_{np}}) \Big] .
$$

Analyzing individual components, again we obtain

F^2 - «Voluntary» abatement points in the direction of greater density. (Higher fines increase density, allowing R to go up, and land prices to drop. This makes abatement increase and R fall). The emission-residual concentration effect is negligible; there are two opposite signs.

F^3 - The term $(q_R R_{NP} NP_p)\dfrac{2r\pi}{q^2}$ makes d increase. (Higher fees make land prices fall, abatement increase, R drop and with it the size of the desired lot, and density d go up. On the other hand, the increased price p has the opposite effect again via the demand for land and goods.) The net effect again is most likely nagligible in quantity.

F^4 - This term definitely increases density.

F^5 - As fees increase, the demand for NP becomes higher and R drops, which makes land prices go up (and hence d increases). Better environmental quality makes land users satisfied with smaller lots (d increases also). The net effect is positive.

Density effects in the goods (F^4) and antipollution device (F^5) markets are positive. The effects in the other components of the model (environment and land market) are ambiguous. It should be mentioned, however, that there are more terms pointing in the direction of higher densities towards the center than in the case of no market for abatement devices (see Schubert, 1979).

Summing up, it seems plausible to argue that density in the urban core regions will increase (note that $|J|$ is positive for small r) and the density in the urban ring will decrease, a process often referred to as «reurbanization» (van den Berg *et al.*, 1981).

6. Conclusions

The model presented demonstrates that an urban system responds to changes in (exogenous) policy variables in a very complex way. Effects are usually contradictory in direction and size, and the net outcome can only be determined by means of empirical analysis. Unhappily all the necessary data for such an enterprise are not available. It seems plausible, however, that if the environmental policy outlined in the preceding section (fines or fees on excessive emissions) were employed, there would be a tendency toward a more compact city which could outweigh the «disurbanization» effects.

A serious drawback of the model used for this analysis is its being static. Cities and systems of cities seem to follow life cycles of growth and decline and concentration and dispersion (van den Berg *et al.*, 1981). In each stage of development there appear to be characteristic constellations of the relevant variables, a fact that makes comparative statics only relevant within narrow margins.

Aknowledgements

I want to thank P. Korcelli and G. Leonardi of IIASA for their comments on an earlier version of this paper and P. Fritz for his art work.

References

Abramovitz M., Stegun I. (1972) *Handbook of Mathematical Functions*, Dover, N.Y..

Alonso W. (1964) *Location and Land Use*, Harvard University Press, Cambridge.

Anderson R., Crocker T. (1971) Air Pollution and Residential Property Values, *Urban Studies, 8*, 171-180.

van den Berg L., Drewett R., Klaassen L., Rossi A., Vijverberg K. (1982) *Urban Europe: A Study of Growth and Decline*, Pergamon Press, Oxford.

Bökemann D. (1977) Theorie der Standortproduktion, Discussion Paper 11, Institut für Stadt-und Regionalforschung, Technische Universität Wien.

Edel M. (1972) Planning, Market or Welfare? Recent Land Use Conflicts in American Cities, in Rothenberg J., Edel M. (eds.) *Readings in Urban Economics*, MacMillan, N.Y., 134-151.

Feller W. (1980) *An Introduction to Probability Theory and its Applications*, Wiley, N.Y.-London-Sidney.

Fisz M. (1973) *Wahrscheinlichkeitsrechnung und Mathematische Statistik*, VEB, Deutscher Verlag der Wissenschaften, Berlin, 132-137.

Isard W., Liossatos P. (1978) *Spatial Dynamics and Optimal Space-Time Development*, North Holland, N.Y..

Muth R. (1969) *Cities and Housing*, University of Chicago Press, Chicago.

Richardson H. (1978) *Urban Economics*, Dryden Press, Hinsdale.

Ruff L. (1972) The Economic Common Sense of Pollution, *The Public Interest, 19.*

Schubert U. (1973) Instrumente des Umweltschutzes, Institute for Advanced Studies, Vienna.

Schubert U. (1979) Environmental Quality and Urban Land Markets, Ph. D. dissertation, University of California, San Diego, (University Microfilms: 29-72, Anne Arbor).

von Thünen J. (1826) *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*, Hamburg.

Résumé. Le bût de cet essai est de construire un modèle d'équations simultanées d'une économie urbaine permettant d'évaluer les impacts des politiques pour l'environnement sur le procéssus de sub-urbanisation et ré-urbanisation. La variable qui montre ces éffets est une mesure de la densité d'utilisation du sol. Le modèle est composé éssentiellement de quatre éléments - un marché du sol urbain, un marché des biens composés, une déscription d'un procéssus d'émission de résidus en tant que sousproduits nécéssaires d'activités qui utilisent le sol (production, consommation, déplacement), une déscription des éffets cumulatifs et diffusifs de ces résidus par rapport à l'environnement (p. ex., pollution de l'air et dômage causé par le bruit).
Les politiques pour l'environnement sont ensuite introduites dans le modèle d'équations simultanées (promulgation de normes pour la sauvegarde de l'environnement, ordonnances pour l'installation d'équipements de contrôle, taxations, etc.). On discute ensuite l'hypothèse que si d'une part une politique anti-pollution éfficace pourrait engendrer une tendence à la ré-urbanisation, de l'autre l'aire propre pourrait induire les utilisateurs du sol à se localiser plus près du centre de la ville.
Les politiques considérées tiennent compte aussi des amendes pour les émissions qui excédent les niveaux acceptables et de la promulgation de normes trop réstrictives qui obligent les utilisateurs du sol à âcheter des équipements de contrôle sur le marché. Cet étude essaie d'évaluer les effets de ces mesures sur la concentration des résidus, le rôle des équipements de contrôle, la densité urbaine basée sur quelques suppositions definies à priori.

Riassunto. Lo scopo di questo articolo è quello di costruire un modello ad equazioni simultanee di un'economia urbana, per mezzo del quale sia possibile valutare gli effetti degli impatti delle politiche per l'ambiente sul processo di suburbanizzazione-riurbanizzazione. La variabile che modellizza questi effetti è una misura della densità dell'uso del suolo. Il modello è composto essenzialmente di quattro componenti: un mercato del suolo urbano; un mercato dei beni compositi; una descrizione del processo di emissione di residui in quanto necessari sottoprodotti di attività che utilizzano il suolo (produzione, consumo, spostamento); una descrizione degli effetti accumulativi e diffusivi di detti residui, con riferimento all'ambiente circostante (per esempio, inquinamento dell'aria e danni causati dal rumore).
Le politiche di salvaguardia ambientale vengono introdotte in questo modello ad equazioni simultanee (emanazione di norme per la tutela ambientale, ordinanze per l'installazione di attrezzature di controllo, tassazione ecc.). Segue quindi una discussione sull'ipotesi che, se da un lato una politica antinquinamento efficace potrebbe generare una tendenza verso la riurbanizzazione, da un altro lato l'aria pulita potrebbe indurre gli utilizzatori del suolo a rilocalizzarsi più vicino al centro della città.
Le politiche considerate includono multe per le emissioni eccedenti fissati livelli restrittivi, le quali inducono gli utilizzatori del suolo ad acquitare gli strumenti di depurazione esistenti sul mercato. Lo studio tenta di valutare gli effetti di queste misure sulla concentrazione dei residui, il ruolo degli strumenti di controllo, la densità urbana basata su alcune assunzioni definite a priori.

Sistemi Urbani, 3, 1982, 453-483 Guida Editori, Napoli, Italia

The planning process: a category-theoretic approach

L. March

Royal College of Art, Kensington Gore, London SW7 2EU, England.

Y.-S. Ho

The Open University, Walton Hall, Milton Keynes, MK7 6AA, England.

Received 30 September 1980, in revised form 8 April 1981

Abstract. The paper reviews alternative types of planning theory and the forms of planning process. It represents a design for a formal planning machine, based on concepts derived from information theory. Category theory and its relevance to the notion of the planning machine is described in some detail. The application of this concept to various planning approaches and institutional frameworks is also discussed.

Key words: planning theory, planning process, formal planning machine, category theory.

The planning process:

«We need a theory and methodology for coping with systems of problems *as systems*, as wholes, as indivisible sets of interdependent elements. These are precisely what the methodologists of *planning*, in contrast to the methodologists of problem solving, try to provide. Planning is essentially an effort to deal with sets of interrelated problems in a way that takes all their interrelationships into account. We do not have as well developed a methodology of planning as we do of problem solving. We have a great need for one» (Ackoff, 1974):

a category-theoretic approach:

«There is some justice, in S. Papert's observation that MacLane's categories are a device for laying hold of mathematical operations rather than of 'mathematics itself': they constitute yet another example of that reflective abstraction which derives its substance, not from objects, but from operations performed upon objects... These facts have an important bearing both on the nature and on the manner of construction of models». (Piaget, 1971).

«Models which have been developed for social phenomena, but are without any roots in the real world, can be useful for the development of social theory... If (the social scientists) holds in abeyance his questions about the model's applicability, and concentrates instead on the formal mechanism of the model, and on how this relates to the substantive postulates, he may learn much from such models. Since they are unconstrained by any correspondence to actual phenomena, they can exhibit many more variations in structure, and many more types of mathematics, than a similar set of models constructed to conform to actual or idealized social phenomena» (Coleman, 1964).

1. Positive and normative social and planning theories

Britton Harris (1978) has proposed a «fourway contingency table» in which to examine «the positive and normative aspects of social and planning theory». From our point of view this is a useful tabulation which is in agreement with our own approach. However, we interpret the four areas independently from Harris. First we define our principle terms:

1.1. *positive planning theory* is a body of organized knowledge describing what planners do and how they do it. Typically, planning theory of this kind is developed inductively through empirical studies and surveys of historial events and contemporary cases;

1.2. *normative planning theory* attempts to propose what planners ought to do and how they should go about doing it. In Harris's terms «in any given context, planning has to solve certain problems. The analysis of planning is possible, one which is not limited to historically observed planning activity nor to the behaviour of planners and designers and their organizations; such an analysis could follow a much more deductive approach». Normative planning theory is required both to criticise existing planning systems, processes and techniques, and to design new systems, processes and techniques;

2.1. *positive social theory* assumes that society is susceptible of analysis in purely objective mechanistic terms and that social values and normative standards are mere epiphenomena. Such an approach to social science may be characterised at its best by Vilfredo Pareto's work. In recent years some planners have made use of mechanistic models of human behaviour concerning matters of relevance such as migration, location and traffic;

2.2. *normative social theory* is evaluative and imperative. In its evaluative mode the theory is critical, and in its imperative form it may be ideological or utopian. The normative sciences are traditionally identified with ethics, aesthetics and logic. Assumptions concerning welfare, visual quality and economic rationality, for example, reflect these interests in planning practice. On the one hand plans and proposals which set out specific goals (imperatives) may be classed as normative constructs; and on the other hand the tools and techniques used to evaluate planning actions also form a normative system. What *ought* to constitute a plan is one thing, and how that plan *ought* to be determined is another. The former is an ideological problem, and the latter is largely operational (see Coleman, 1964).

2. A model of planning processes

The model we propose places the planning system within the social environment as a whole. It attempts to provide a framework within which the «facts» which inform the planning system are related to «values» and «interpretations». The «facts» we consider are essentially those external to the planning system itself which nevertheless form the system of interest for planning (e.g. facts concerning the state of the urban system). The «values» and «interpretations» are of two kinds: those determined externally between the planning system and its environment, and those determined within the planning system. Our model overall is normative, but it contains a positive, descriptive component. The model is employed primarily to develop a unifying metatheory of planning in order to provide a conceptual framework within which specific studies may be related.

The proposed model, being normative in purpose, is not concerned with the observed *practice* of planners in this or that situation either with respect to institutional organisation or to decision-making. In this respect the model is one of «competence» rather than «performance» (see section 7.4.).

3. Planning as an information processing system

The model views planning, like design, as an *information processing system* (Simon, 1969; Simon, Newall 1972). Classically, such a system comprises four parts: *receptors*, *effectors*, a *processor* and a *memory* (see fig. 3.1). Such a model is familiar enough today, but it derives

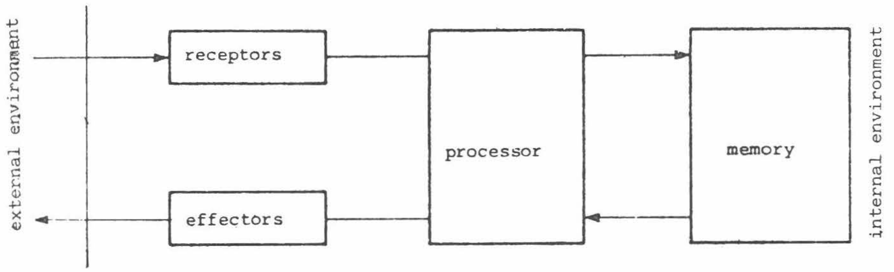


Figure 3.1 General structure of an information system

originally from Craik's (1943) speculations on *The Nature of Explanation.* A purposive system according to Craik involves

«(1) 'translation' of external process into words, numbers and other symbols;

(2) arrival at other symbols by a process of 'reasoning', deduction, inference, etc., and

(3) 'retranslation' of these symbols into external processes (as in building a bridge to a design) or at least recognition of the correspondence between these symbols and external events (as in realising that a prediction is fulfilled)».

The «translation» and «retranslation» is achieved by the encoding and decoding processes of the receptors and effectors acting as transducers and preprocessors. The transformation of symbolic matter in stage (2) requires both a processor and memory. The important point of Craik's approach is that an individual person acts upon the world through the mediation of a *model* of the world. Such a model is conceived as a *machine* and it may be correctly said that Craik's work provides the overture to mechanistic modelling in *artificial intelligence*, *cognitive psychology* and modern *linguistic studies*. The explicit use of Craik's concept is acknowledged by Stiny, Gips (1978) in *Algorithmic Aesthetics: Computer Models for Criticism and Design in the Arts*. These authors set out in general terms the algorithmic constituents for a criticism/design machine (*).

The aim of this paper is to carry out parallel investigations to specify the design of a *planning machine* in general but nevertheless precise terms. Dunn (1974) has characterized such a machine in the context of his Resources for the Future study of social information processing. His model (fig. 1, p. 28) is essentially the same as the Simon diagram above. Dunn also points out that individual human information processing is a necessary metaphor for a social system: in this respect both Craik's psychological model and Simon's human problem-solving system are useful reference points. However, as M. Batty (1974, 1975) and S.E. Batty (1977) have demonstrated, a planning machine must reflect a *plurality* of interests including conflicts between parties to a plan. The similarities and differences between *human* information processing and *social* information processing are well summarised by Dunn (pp. 29-46). The unique characteristics of social information processing include:

(1) the absence of a superordinate «self» in a social organization;

(2) the political dimension of conflict regulation;

(3) the artificial replication of social information processing (such as exosomatic technological evolution and amplification of natural human processes);

(4) the explicit generation of symbolic representations between the machine and its environment and within the machine itself.

(*) Since this paper was presented Stiny, March (1981) have developed this concept further in a general design context.

These features must be incorporated in the design of the proposed planning machine.

It is the fourth of Dunn's characteristics - *symbolic representation* - that needs our most careful theoretical attention. Simon, Newall (1972) identifie the representation problem as fundamental in problem-solving. Most progress, they remark (p. 90), has been made in developing problem-solving *methods* within a given representation. However,

> «... much less has been learned about the nature of representations, and their selection and modification».

Dunn suggests that social information processing has two representational tasks related

«(1) to the field of observation formed by orders of system hierarchy relative to an observer (or references) system, and

(2) to the levels of information processing required by the adaptive situation».

The two tasks are further subdivided by Dunn. From the observer's (processor's) point of view there is an hierarchy of systems and corresponding representation problems:

(1) the processor itself: self-representation;

(2) the processing system as a whole including the relationships of the processor to other components such as the receptors, effectors and memory: representation of the *internal* operating system;

(3) the system of interest: representation of the *external* operating system; and

(4) the wider environment: representation of the *peripheral* system.

There is also an hierarchy of information processing:

(1) the organisation of descriptive data: the representation, for example, of simple inventories and files;

(2) the organisation of relational or activity data: the representation of elementary structures;

(3) the programmatic systematisation of data: the representation of systems and algorithms;

(4) the learning process: the representation of theories and heuristics.

This hierarchy follows a pattern of increasing «consciousness» within the information processing system, and Dunn draws attention to an analogy with Piaget's developmental stages for the individual in terms of social development. Formally, Dunn's four levels of information processing correspond to the representation of

(1) the entities forming the data sets (*objects*)

(2) relations between objects (*structures*) (*)
(3) mappings between structures (*systems*) (**), and
(4) evaluation and selection of systems (*theories*) (***).

Thus Dunn suggests a sixteenway table for the representation problem;

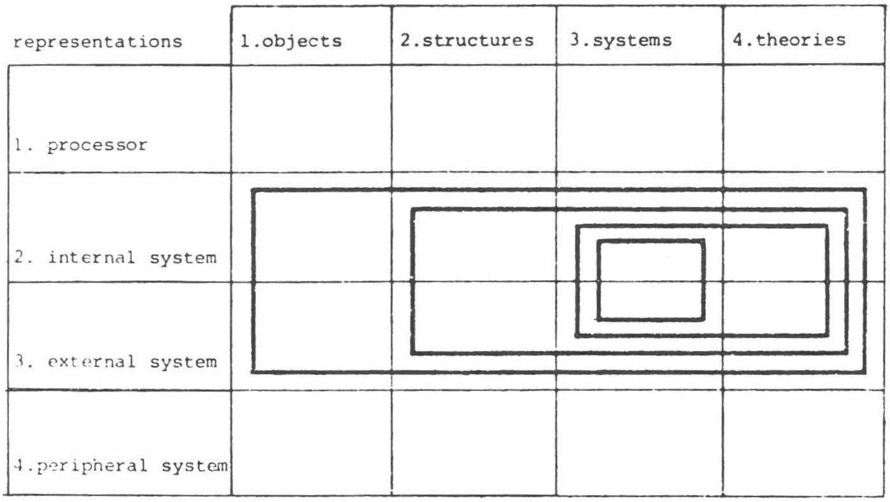


Figure 3.2 The variety of entity representations

Our own interests are indicated by the boxes in fig. 3.2. We shall not be concerned with the problems of representation within the processor itself (perhaps an artificial intelligence/linguistics/computer science problem), or with the peripheral system (a problem for social science and theory in general). We *shall* be interested in representing the *empirical states* of the external system, and in representing the *symbolic forms* within the internal system. Further, in order of interest, we shall concentrate on the representation of structures, but mostly of systems and theories.

(*) This notion of *structure* is to be found in Harary, Norman, Cartwright (1965).
(**) This classicial definition of *system* follows Hall, Fagen (1956).
(***) This view of *theory* formation arises from Lofgren's (1972) learning hypothesis together with Solomonoff's (1964) and Chaitin's (1966, 1975) discussion of scientific theory in terms of algorithmic parsimony.

It should not be construed that we believe in the existence of *ideal* representations. Our approach assumes that there are *many* representations for a given situation, but that an optimal representation may be sought for particular purposes. Indeed, in part, the representations will be chosen in regard to the *transformations* which take place within the information processing system. The nature of these transformations constitutes our second matter of theoretical concern.

4. Preliminary design of a formal planning machine

«When the mapping of actions on states of the world is problematic, then, and only then, are we faced with genuine problems of design. Hence, if we wish to construct a formalism for describing design processes, that formalism will need to contain at least two distinct sublanguages: a sublanguage for actions and a sublanguage for states of the world. To the extent that these two sublanguages are related simply to each other, so that an expression in the one can be translated without difficulty into the other, the problem of design will vanish. For the problem of design *is* the problem of making that translation» (Simon, 1967).

1. The planning process is depicted as a system within a sociopolitical environment.
2. The inputs comprise observations of the state of the environment, and the outputs are proposals and plans to guide actions in the environment. The planning system may be divided into four subsystems: the *receptor*, the *processor*, *memory* and the *effector*.
3. For the purposes of the model proposed here it is assumed that the receptor and effector embody encoding and decoding algorithms. The processor and memory which characterize the planning process itself, are wholly concerned with the storage and transformations of coded data sets.
4. The processor consists of two interacting subsystems: one a positive system carrying descriptions about material states, and the other a normative system holding information about decisions (see, for example, Simon, 1967).
5. Both subsystems depict their respective spaces at appropriate levels, and both include selection procedures whereby a particular material situation or a particular decision may be referred to and isolated.
6. The normative decision subsystem may be separated into two parts. Some values are enshrined in law reflecting the norms of a given sociopolitical system: such values are said to be *prescriptive*. Other values represent the views of individuals and groups: these are said to be *elective*.

7. Since in Western society it is assumed that elective decisions are arrived at through market mechanisms, and that prescriptive decisions are largely those imposed by the State and its agencies, it seems reasonable to ascribe the one to economics and the other to politics. The material subsystem is essentially technological in dimension. The expanded model given in 6 above thus has three aspects (see fig. 4.1).

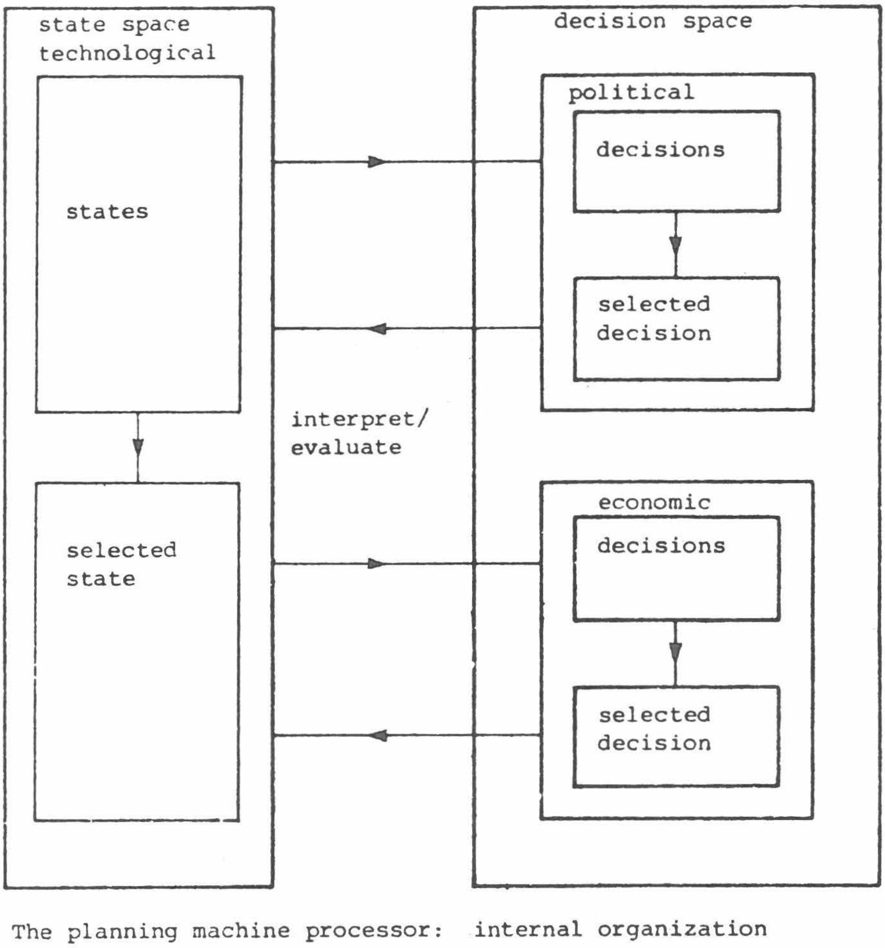


Figure 4.1 The planning machine processor: internal organization

8. The political aspect is dependent in part on normative social theory in terms of national goals and in terms of protocol. Normative social theory enters out model only as norms to be satisfied and procedures to be followed. Political changes *per se* are outside the model. The economic aspect of the model is dependent on normative

theories of rational decision making. The technological aspect contains descriptions of social behaviour derived from positive social studies. All subsystems carry optimizing or «satisficing» techniques for isolating particular situations for evaluation derived from normative planning theory. As we have said before the whole model owes little to positive planning theory (the way planning is actually conducted in practice).

9. The memory of the planning machine stores its history: its «experiences» of «facts», «interpretations» and «values» (*). It holds both experiences of the external world *and* of the internal operating system. Such «experiences» are called upon from time to time as required by the processor for specific tasks. The more economical way of storing experiential data is in the generative mode whereby data are reconstituted as demanded. The least economical way of storing data is item by item. The former mode requires that the data set be defined *intensionally*,

$$\{ x \mid x \text{ is generated by algorithm } A \} ,$$

while the latter defines the data set *extensionally*,

$$\{ a_1, a_2, a_3, \ldots a_n \} \; (**) \; .$$

5. The mathematical treatment of the processor

In detail the processor we propose comprises a description of the *state space* and a description of the *decision space*. Both spaces are described mathematically. We assume that the data is encoded in discrete form and that it is organised in well defined, finite sets. Formal processing of these sets requires *closure* (in the mathematical

(*) For a brief discussion of the need to appreciate that «'facts' are inseparable from 'interpretations', which in turn are determined by 'values'» in history; and that interpretations are dependent on theoretical concept formation in regard to «specific structures which have specific histories» see Stedman Jones (1973).

(**) Solomonoff (1964) assumes that an experiential set such as a scientist's observations may be encoded as a binary series. A *theory* can be regarded as an algorithm capable of generating the series and extending it, that is, predicting potential «experiences». For any given set of «experiences» there are always several competing theories, and a choice must be made between them. Solomonoff suggests that the smallest algorithm, the one consisting of the fewest bits, be selected. The smaller the program, the more comprehensive the theory and the greater the degree of understanding. As Knuth (1973) points out

«It has often been said that a person doesn't really understand something until he teaches it to someone else. Actually a person doesn't really understand something until he can teach it to a computer, i.e., express it as an algorithm... The attempt to formalize things as an algorithm leads to a much deeper understanding than if we simply try to understand things in the traditional way».

sense) and suitable closure operators must be constructed. This is the first essential step in modelling. Without mathematical closure there can be no constructive development of algebraic or topological structure, and hence no possibility of establishing an homomorphism between structures in the real world and abstract structures in the processor.

Many different mathematical structures could be employed (Piaget, 1971). One useful classification depends on the type of *relation* used to structure the data set. Effectively there are three broad classes of relation (Schreider, 1975):

equivalence,
order, and
tolerance.

The equivalence relation has the effect of partitioning the data into disjoint sets to create a nucleated structure. There are various kinds of order relations leading to treelike (hierarchic) or lattice-like (reticulated) structures. The tolerance relation does not require that sets be equivalent, but that sets merely *resemble* one another. With tolerance, the data set is not nucleated as with equivalence, but forms a system of intersecting subsets capable of preserving organizational complexity. One representation of such a system is known as a *simplicial complex* (and is the subject of considerable research in the social science context by Atkin (1974, 1977) and his coworkers). The work of a group of French writers provides a useful introduction to these mathematical ideas as they relate to the social sciences (Centre de Mathématique Sociale, Ecole des Hautes Etudes en Sciences Sociales, 1976).

A *structural representation* is a data set and a collection of relations defined on it. An *ensemble* consists of a collection of *representations*. The ensemble is structured - is a *system* - if relations are defined between the representations. Other ensembles may be defined constructively at different levels and again relations established between the levels. In particular, we have in mind the construction of power sets. The ensembles may comprise representations of actual situations, or by construction all possible situations. The latter form of ensemble is the source of both planning *forecasts* and the generation of planning *options*. These two kinds of ensemble allow for two distinct forms of implication: the indicative «*... since ... then*» and the speculative «*... if ... then*».

The state and decision spaces are conceived as comprising representations with relations defined between them. The spaces themselves are multilevel, hierarchical structures. There are mappings between one level and another. There are also mappings between the two distinct spaces. Without characterizing the precise nature of the objects, relations or mappings we may still develop a powerful mathematical theory of the planning process (as depicted by the processor). This is done using the concepts of category theory.

Category theory talks about categories comprising *objects* and *arrows*, and mappings between categories defined by *functors*. Some elementary details of our proposed use of category theory are given below, but suffice it to say here that the spaces in our model may be characterized by categories. The objects are the representations and the ensembles at various levels. Each class of object (representations, ensembles) belongs to a category with appropriate arrows defined between objects within each category. Between categories functors are defined. Some functors operate vertically *within* a given space, and others operate horizontally *between* the state space and the decision space. Arrows structure each level: «vertical» functors provide structure in a given space as a whole, and «horizontal» functors make it possible for the material and value systems to interact. [It is of course possible to have more than one set of objects and arrows at any given level (i.e. there may be several categories at one level) in which case there could be «horizontal» functors within a space. For simplicity we shall ignore these in the present discussion in order to preserve the distinction we have already made above].

Attention may now be drawn to the fact that we are able to discuss various levels of planning activity from detailed design to strategic policy. The processor, may be used in a creative (generative) mode or a critical mode, and it may be used to exercise planning control.

6. Some aspects of category theory and its relevance

It is not our intention in this note to dwell on the technicalities of category theory. The advanced mathematical theory is well set out by MacLane (1972). A useful summary is given in Howson (1972) which we have reproduced as appendix to this note for reference. For system scientists working in a wide variety of fields the main concepts have been described by Arbib, Manes (1975):

«What, then, is the *categorical imperative* - the set of core concepts of category theory which should be shared by this diverse audience before they pursue more specialized avenues tailored to their own area of interest? Our answer is threefold. First is the ability to think with *arrows*: to express key concepts in terms of mappings (we call them *morphisms* in the general setting of category theory) rather than in terms of set elements. Second is the realization that collections of mathematical *structures* find convenient characterisation in terms of arrows. Third is the use of *functors* as the appropriate tools with which to compare different domains of mathematical discourse».

The importance of functors lies in their modelling properties. Given an observed social situation it is a task of a planning machine (information processing system) to model these entities and relationships through some formal, symbolic system (Dunn, 1974).

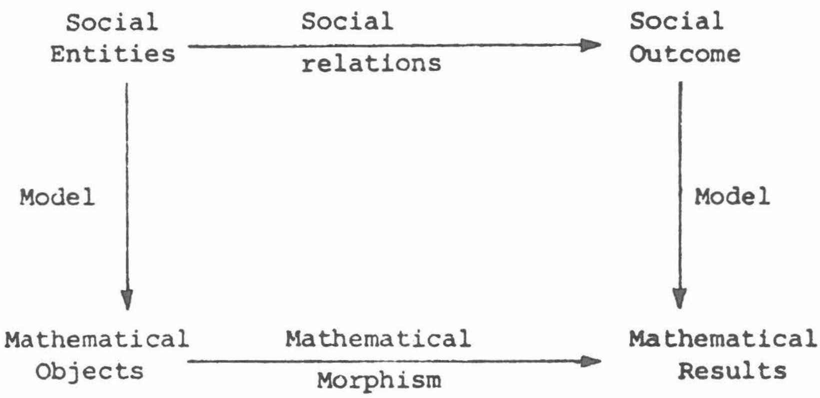


Figure 6.1 Mathematical modelling of a social situation

In practice, the usefulness of the model will depend largely upon whether we can go from the mathematical result to the social outcome. That is, the usefulness depends on the possibility of reversing the righthand arrow and on finding a suitable interpretation for translating results in the model back to the outcomes in the social situation. The reversal of the arrow corresponds to the finding of the *inverse morphism* (in particular an *invertible functor*, see Appendix). In addition to transformations from the social world to the symbolic world of the planning machine, there are also transformations within the machine between different symbolic structures. That is to say, having set up a formalism in one particular way, we may find ourselves wanting to tackle the problem from an entirely different standpoint, because perhaps our initial statement of the problem raises procedural or computational difficulties. Category theory provides the proper framework for discussing these transformations.

A *category* C has objects A, B, C,..., and *arrows* f, g, h,... (or morphisms of objects):

$$A \xrightarrow{f} B, \quad B \xrightarrow{g} C$$

and so on. If we consider the state (or decision) space in the planning machine as categories, then the objects will represent the states (or decisions) and the arrows will be morphisms transforming one state (or decision) into another.

Example. In linear systems such as Batty's Markovian design machine (1974) the category is a vector space over a Borel field, the objects are vectors and the arrows are stochastic transition matrices.

The arrows in a category may be composed so that two successive arrows

$$A \xrightarrow{f} B \xrightarrow{g} C$$

may be compounded into one arrow

$$A \xrightarrow{gf} C \ .$$

Arrows are associative so that (hg) f = h (gf), and each object A has an identity arrow, that is

$$A \xrightarrow{i_A} A \ .$$

It is noted that the transition matrices in the example above satisfy these composition laws.

A diagram in a category is a *directed graph.*

Example. A poset is a category (Arbib, Manes, 1975). Consider the poset of four elements $p \leq q \leq r \leq s$. The diagram of this category has four vertices labelled by the objects p, q, r, s. The labelled edges of the directed graph are the arrows of the category. Thus $f_{xy} : x \rightarrow y$ if $x \leq y$, otherwise if $x \nleq y$ there is no arrow.

For f_{xx} we write i_x, the identity morphism, and this is a loop at x in the diagram.

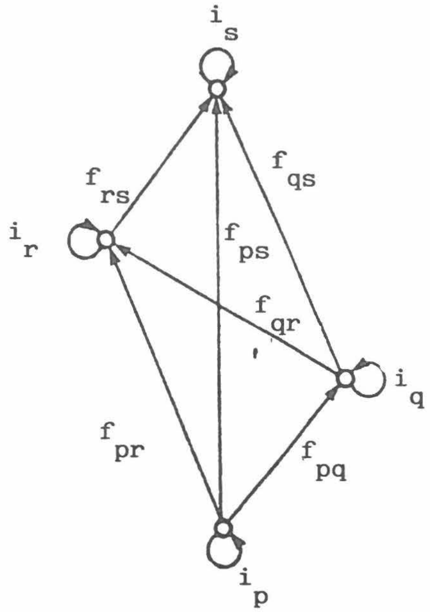


Figure 6.2 A categorical diagram for a poset

Closely related to the concept of category is the concept of functor. Functors are morphism transforming one category into another. A functor H from a category *C* to another category *D* is a morphism mapping the objects in *C* into objects in *D*. For each arrow $f: A \rightarrow B$ in *C*, the functor H gives an arrow $H(f): H(A) \rightarrow H(B)$ in *D*:

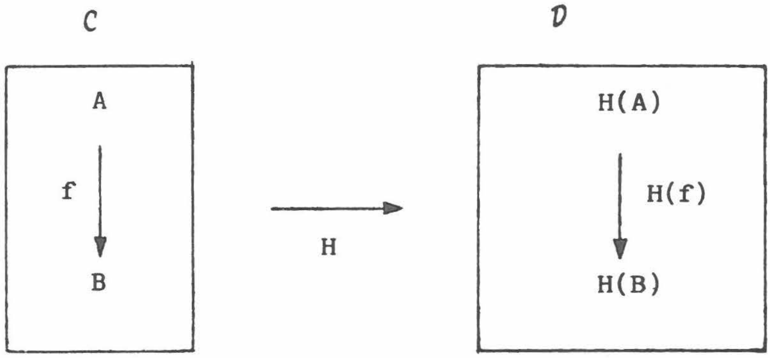


Figure 6.3 A mapping of the functor $H: C \rightarrow D$

The functor has to satisfy the two conservation laws:

1) conservation of identity
   $H(i_A) = i_{H(A)}$, and
2) conservation of composition
   $H(fg) = H(f)H(g)$.

Diagrams are used to present category-theoretic relationships. We may represent the action of the covariant functor above by a commutative diagram:

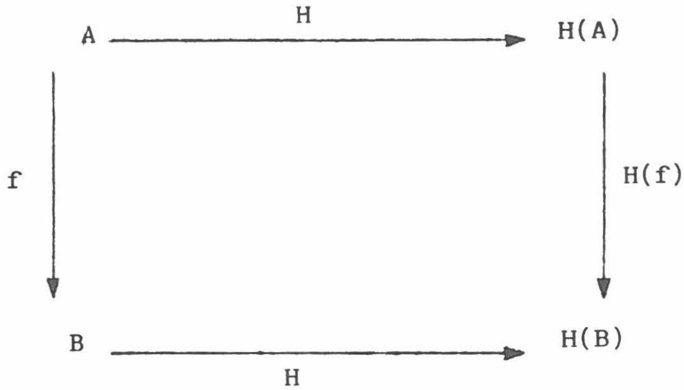


Figure 6.4 Diagrammatic representation of the laws of functor

in which $Hf = H(f). H$.

Example. Consider a category C as domain and a category D as codomain. Each category will have a structure diagrammed by a directed graph of arrows between its objects. For clarity this structure is omitted in the illustration below. There are four morphism types depending on the number of objects in C and D respectively and on the pattern of functors.

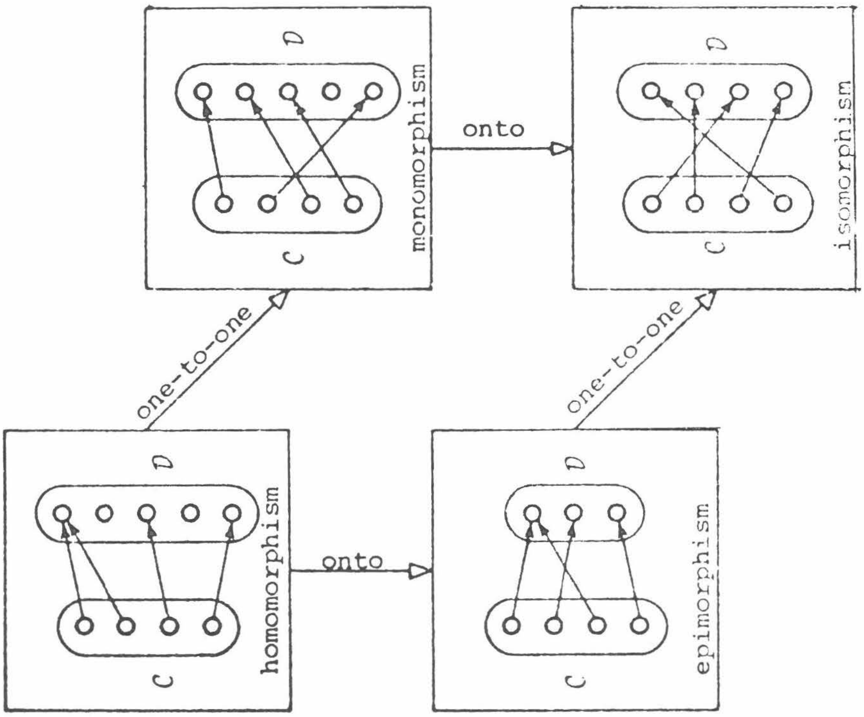


Figure 6.5 A classification of morphism types

Example. Atkin (1974, 1977) employs the category of simplicial complexes S, the category of chain complexes C, and the category of Abelian groups A to describe the complex organization of social situations and their environments. The transformation from one category to another is carried out using homology groups and induced homomorphisms as functors. The transformation means that it is possible to move from a topological representation of the social situation to an algebraic representation. Each representation provides a distinctive «picture» of the social scene, and a different data structure (see Kuratowski, 1972).

One particular functor is known as the *forgetful functor* on account of its ability to forget structure in mapping from one category to another. It is clear that many planning models implicitly use forgetful functors in mapping from the complex structure of observed behaviour to a simplifying model of that behaviour. In particular, few planning models preserve the topological structure so apparent in social situations and affairs.

Typical categories include the equality-preserving systems from monoids to linear algebras used by social scientists and planners to describe the states of social situations and environments (e.g. Piaget, 1971); the order-preserving systems from posets to boolean lattices used by value theorists and logicians (e.g. Burks, 1977); and the tolerance-preserving systems which since Zeeman, Buneman (1968) (*) have been shown to apply to the topology of social situations (e.g. Atkin, 1974, 1977). All of these categories are described in the Centre de Mathématique Sociale, Ecole des Hautes Etudes en Sciences Sociales (1976), publication.

In terms of the planning process the mapping between states and decisions is a functor between the categories representing states and decisions respectively. The interdependence of technological possibilities within the state space and the shifts in opinion within the decision space is property reflected in the commutative diagram.

The essential property of functors is their *structure-preserving* capability. They allow us to change from a representation of a social system by one mathematical system (category) to another representation using a different category, and they make it possible to relate a category describing a situation in the state space of a planning machine to another category in the decision space. The functors in the latter case are *evaluators* and *interpreters*, while in the former instance they are *analysers* and *synthesisers* permitting us to move between levels in the hierarchy of systems represented in the state space (for example: between characteristics and designs, or designs and plans, or between plans and policies in the state space; and between individual preferences and community desires, or community wishes and national priorities in the decision space).

We believe that the theory of categories and functors provides a convenient framework in which the formalism of planning can be organized with some success. Category theory arose when

(*) It is noteworthy that Zeeman uses examples of environmental perception in his pioneering work on tolerance spaces and the brain.

mathematicians began an investigation of *natural correspondences* between mathematical systems about thirty years ago. Since then the theory is appearing with ever increasing frequency in many application areas including the theory of automata, linguistics and control theory (Bobrow, Arbib, 1974; Manes, 1973), and an interesting case is made out by Rosen (1978) for its use in the study of natural systems as a unifying framework. The idea of a unifying framework for the physical sciences has been recently explored by Tonti (1975) using the methods of combinatorial topology and associated algebras. This work is close in spirit to the multidimensional systems approach of Kron (1963) which we have noted earlier. Tonti's work does not use category theory explicitly, but as Kuratowski (1972) shows the relationship between algebraic topology and its operations and categories and functors is an intimate one. We would hope to use our category-theoretic approach in the design of a planning machine to show the relationships between the many types of planning models now available.

7. The planning machine: some interpretations and relation to other works

7.1. *Actions, decisions and logics for design and planning*

Simon (1967) argues that a problem of design [*a fortiori* creative planning] exists when

«(1) there is a language for naming actions and a language for naming states of the world,

(2) there is a need to find an action that will produce a specified state of the world or a specified change in the state of the world, and

(3) there is no non-trivial process for translating changes in the state of the world into their corresponding actions».

He then goes on to discuss decision situations treated by classicial economics and statistical decision theory: «the problem is considerably simplified», he writes,

«(1) because the set of possible actions generally consists simply of the set of vectors of the known command variables;

(2) the language of actions (i.e. of the command variables) is essentially homogeneous with the language of states of the world; and

(3) states of the world are assumed to be connected with the command variables by well-defined, and relatively simple functions.

Under these simplying conditions, algorithms often exist that are guaranteed to find, with any reasonable amounts of computation, actions corresponding to desired states of the world. Thus the classical maximizing algorithms of the calculus combined with the newer

computational algorithms of linear programming and similar formalisms are powerful enough to handle quite large problems when the prerequisite conditions are met (*).
...I have called them problems of «choice» rather than problems of «design» because the set of possible actions is given, in a very clear sense, and because the determination of the correct action involves a straightfoward, if somewhat tedious, deduction from the description of possible states of the world».

The same objections might be applied to entropy-maximizing decision models (Tribus, 1969). Although, for example, Tribus distinguishes {data, theory} in the descriptive *state* space from {actions, outcomes, utilities} in the *decision* space, mathematically the two spaces remain homogeneous. Further, the set of possible actions must be well-defined and exhaustive. Failure to satisfy these requirements «represents an inability to be definite about one's aims». It should be noted that Tribus's spaces are not *structured* in the sense we have described earlier. The components of Tribus's model are essentially correct (see March, 1975), but the model itself is deficient in the same way that other classicial decision models fall short.

In March (1975) three modes of reasoning are identified with the design and planning activity. These three modes are related to C.S. Peirce's system

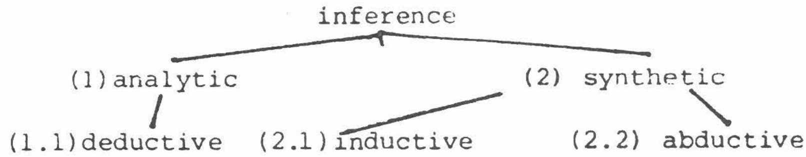


Figure 7.1 Peirce's classification of inferences

We may consider a universe of *objects* possessing certain *characteristics* (**). There are *rules* which determine which objects possess which characteristics. The following modes of reasoning are desirable:

(1.1) deduction: given a *rule* and an *object* determine a *characteristic* possessed by the object;

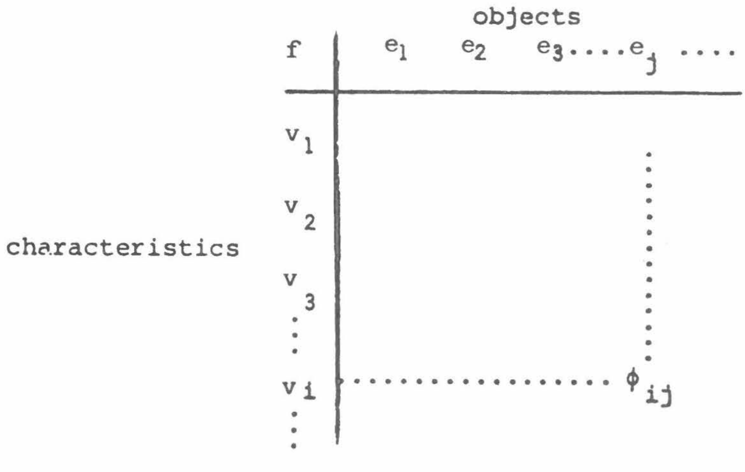
(*) In planning many of these programming techniques were usefully described in Scott (1971). More recently the technique of entropy-maximizing has been related to geometric programming. Simon's discussion holds for this whole class of programming methods.
(**) This approach was adopted by Lancaster (1966, 1971) in consumer demand theory. It appears in philosophical discussion in Leonard's (1957) development of Russell's wellknown theory of types.

(2.1) inductive: given an *object* and a *characteristic* determine a *rule* relating the two;

(2.2) abduction: given a *characteristic* and a *rule* determine an *object* possessing the characteristic.

Whereas the deductive mode of reasoning is determinate, the inductive and abductive are not. Peirce suggests that induction is evaluative, and abduction is creative (*) and these ideas are developed in March (1975), while in March, Batty (1975) the three modes are related to the Bayesian framework of Tribus's design model.

The three modes may be expressed by the triple (V, f, E) where V and E are disjoint sets and $f: E \rightarrow P(V)$, the power set of V. If V is the set of *characteristics*, then P(V) represents all possible complete characterisations of *objects* in the universe. E is the set of objects in the universe and f maps these into equivalence classes defined by characteristic similarities. If objects e_1 and e_2 are both completely characterised by $w \in P(v)$, then e_1 and e_2 belong to the same equivalence class. This system (Graver, Watkins, 1977) may be expressed as a zero-one matrix:



where $\Phi_{ij} = 1$ if object e_j possesses characteristic v_i and is 0 otherwise. If f, the rule is given then there is a precise answer to the question: does object e_j possess *characteristic* v_i or not? This is the analytical deductive mode of reasoning. If the rule f and a characteristic v_i are given there may or may not be an *object* e_i with the characteristic and

(*) See Simon's (1977) support for the abductive mode in *Models of Discovery.*

indeed there may exist several such objects. This is the synthetic abductive mode of reasoning. The object(s) - in this context - may be thought of as a *design(s)* or *plan(s)* satisfying a given set of requirements (characteristics). Finally, if an object e_1 is given with characteristic v_i then it is clearly problematic to determine the *rule* f. This is the inductive problem at its crudest - generalising from one particular instance.

The model above may be developed to include the analysis of an object into a complete characterization, to include the synthesis of a set of objects possessing a given set of characteristics, and to include an attempt to determine a synthesis or general rules given the complete characterisations of a representative set of objects. The last problem might be solved by finding an algorithm to determine the zero-one entries of the matrix, and this is clearly the task of theory formation as described by Solomonoff (1964). It is our thesis that design or planning requires all three modes of information processing illustrated by this discussion. *Characteristics* might include elementary observations of the environment, the perception of needs and opportunities, specification of actions and objectives. *Objects* might include states, situations, provision of services, plans and proposals, and goals. *Rules* might cover experience, habit, norms, models and theories. The three components are related:

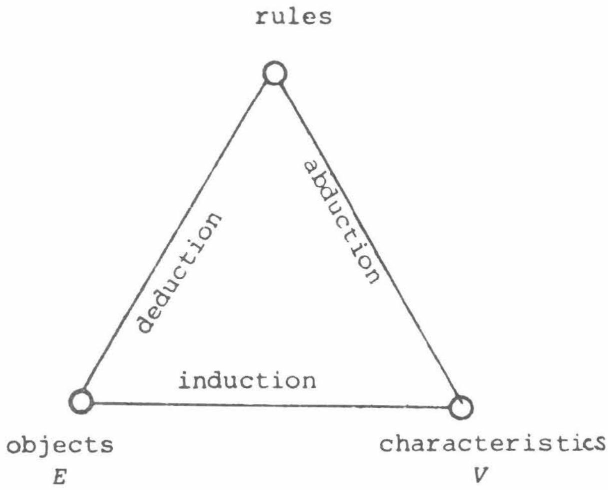


Figure 7.2 Descriptive relations of the three modes of reasoning

The zero-one matrix representation of a theory has been discussed elsewhere, and it should be noted that a theory may also be represented by a *simplicial complex* (Atkin, 1974) or a *bicoloured graph* (Harary, March, Robinson, 1978 and forthcoming). The three representations are isomorphic although for historical reasons the mathematics of combinatorial topology is well-developed commencing with, for example, Aleksandrov, 1956; Pontryagin, 1952) and consequently the simplicial complex representation has certain formal advantages.

The processes within the planning machine are governed by logical argument. However, there may be more than one logic. The choice of an appropriate logic will depend on the task at hand. Simon (1965, 1967) argues that standard propositional and predicate calculi are sufficient for design purposes. He does this by the slightly curious device of using an example from standard optimization procedures which he has argued elsewhere does not correspond to quintessential design (see the quotation cited above 7.1.).

In this situation, Simon sees no good reason for introducing normative, deontic logics for handling «shoulds», «shalls», and «oughts» of various kinds. Also Simon avoids the difficulties of *logic choice* by employing the Wittgensteinian state-description of «all possible worlds» regardless of the ontological questions this ploy raises (see Simon, 1967, p. 62). Despite Simon's conviction, distinctive logics of preference and action have been developed (von Wright, 1962, 1963, 1967) within the same «all possible worlds» construction. Our own interests tend to be concerned with issues such as

(1) the consequences of rejecting the Wittgensteinian universe of «all possible worlds» for a «structured sets» approach involving questions of admissability;

(2) the validity of the rule of *tertium non datur* (excluded middle) particularly in synthetic reasoning;

(3) the weakening of conventional preference postulates to avoid obligatory transitivity and comparability.

For example, in planning we shall not want to insist that *either* project A meets the objectives *or* it does not. A third way might be sought quite reasonably and we would not wish to call this search «illogical». There are logics (e.g. Brouwer's system) which permit us to avoid the exclusion of a middle way. Nor shall we want to insist that *either* plan A is preferred to plan B *or* plan B is preferred to plan A (comparability); that *if* plan A is preferred to plan B, and plan B is preferred to plan C *then* plan A is to be preferred to plan C (transitivity). Again there are calculi which do not require transitivity or comparability (Burks, 1977):

«But at the moment I am interested in the fact that the rule of maximizing utility implies that preferences and uncertainties should be simply-ordered. For I wish to object to the latter, and thereby to the former. More specifically, I shall argue that some uncertainties are not comparable and that some preferences are neither transitive nor comparable».

7.2. *Systems theory and control*

We are familiar with applications for linear systems theory to social systems (March, Ho, 1976). It is to be noted that linear system theory was developed originally to model electrical systems and servomechanisms. The theory views *structures* in the world as isomorphic to *linear graphs*, and *functions* in the world as isomorphic to *linear operators*. Some years ago these limitations were severely criticised by G. Kron (1963) who developed his polyhedral network approach to electromechanical systems using combinatorial topology and tensor algebras. The inclusion of linear graph theory in the wider context of algebraic topology is well developed by Lefschetz (1975). We would not question the successful use of linear systems theory in the social sciences (for example, Stone, 1967), but we are skeptical of its fundamental legitimacy except in a few, limited cases.

Nevertheless, linear systems theory does introduce concepts such as *control*, *regulation*, *change*, *stability*, *limits* and *equilibrium* which belong to the standard vocabulary of social discourse and which any adequate model of planning processes must include. *Time* in social systems is conveniently modelled through linear systems and discussion of *synchrony*, *diachrony*, *statics* and *dynamics* is made possible in an interesting way; however (Cortes, Przeworski, Sprague, 1974):

«The restriction of the study of change to temporal variations enclosed within the bounds of structural invariance constitutes the major limitation of the grammar of linear systems».

A *planning* machine which could not cope with *structural* change would be restricted operationally to incrementalism and piecemeal engineering and this in our view would not be adequate.

7.3. *Planning styles and institutions*

Our planning machine has been designed with a processor consisting of three parts: a positive, *technological* state space; a normative (prescriptive), *political* decision space; and a normative (elective) *economic* decision space. Consider the following possibilities:

	strong	weak
1. technological	T	t
2. political	P	p
3. economic	E	e

where by «strong» we mean that the dimension plays an important part in the planning process, and by «weak» we mean that the dimension does not.

In the developed world we assume that the technological dimension is strong, whereas it is presumed to be weak in the undeveloped world. There are then four planning styles in developed societies represented by the planning machine according to the emphasis given to politics and/or economics. These four styles may be matched to Berry's (1973) and Boguslaw's (1965) typologies of planning modes (see also Couclelis, 1977):

	Berry	Boguslaw
PTE	Incremental progressive compromise. Allocative trend-modifying	Heuristic approach
pTE	Ameliorative problem-solving. Incremental conservation	*Ad hoc* approach
PTe	Developmental leadership. Exploitive opportunity-seeking	Operational unit approach
pTe	Normative goal orientation	Formalist approach

PTE planning is typical of organized capitalism (Gurvitch, 1971). Here the planning system is centralised but its role is essentially *indicative*, acting as a central information processor. Plans are largely carried out through the market economy of organized capital. Both political and economic influences are strong.

pTE planning is typical of liberal democracies believing in *laissez-faire* and non-interventionist government. The free market is allowed its sway.

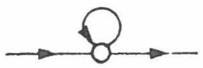
PTe planning is typical of the welfare state and social democracies. The political will is strong and economic forces are tamed (or perhaps ignored).

pTe planning is typical of centralised one-party states. The political and economic dimensions are sublimated in technico-bureaucracy (Gurvitch, 1971).

These four styles of planning are caricatures of course, but they may serve to demonstrate that the proposed planning machine may be able to capture some of the qualitative features of different planning systems.

PtE, ptE, Pte, pte styles might be identified in the undeveloped world, but they all share a «weak» technological base.

So far we have indicated how the machine might reflect particular planning styles, next we suggest how several machines might be organized to create distinct institutional frameworks for planning. Each machine has inputs, outputs and feedback. Let us represent a planning machine as follows



Then we may consider three arrangements of *two* planning machines in series, and three in parallel assuming n-way $(n = 0, 1, 2)$ communications between them (Klir, 1969).

With more planning machines the number of configurations increases rapidly (see fig. 7.3).

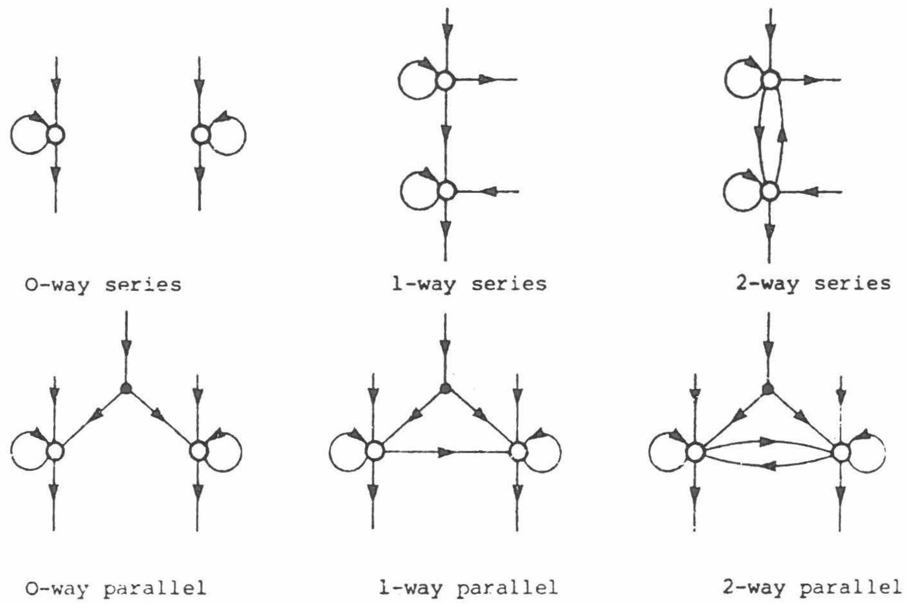


Figure 7.3 Serial and parallel n-way $(n = 0, 1, 2)$ communications between two planning machines

The work of the Batty's (1974, 1977) suggest that some fruitful studies might be made of *systems of planning machines* once the planning machine itself is adequately characterised. This, however, is not our immediate interest and the speculation is left for future investigation. Some work of Gel'fand, Tsetlin (1971) on the behaviour of automata collectives is quite suggestive in this connection.

7.4. *Competence v. performance*

Chomsky (1965) draws attention to a distinction that must be made between studies of «competence» and studies of «performance». In part the distinction is between the syntactic and semantic aspects *and* the pragmatic aspects of an information processing system. The proposal we are making is to investigate planning from the viewpoint of competence (essentially a normative study of planning *constructs*) rather than from the angle of performance (essentially a positive study of planning *practice*).

Interestingly, Chomsky (1963) demonstrates that whereas

«*competence* ... cannot be characterized by a finite automaton... the *performance* must be representable by a finite automaton of some sort».

Whether the same argument holds for planning systems is an interesting point. At this stage we are inclined to believe so. This argument then supports our position of treating the *planning machine* as a theoretical design indirectly related to actual planning procedures. The actual procedures are only observable through effective performance and thus properly belong to a more pragmatic, empirical study.

Our stand is *structuralist* in the analytic sense (Piaget, 1971): «Whereas an empiricist holds to systems of observable relations and interactions which are regarded as self-sufficient. The pecularity of authentic (analytic) structuralism is that it seeks to explain such empirical systems by postulating «deep» structures from which the former are in some manner derivable. Since structures in this sense of the word are ultimately logico-mathematical models of the observed social relations, they do not themselves belong to the world of «fact» [in the vulgar sense]. This means among other things, as Levi-Strauss points out repeatedly, that the individual members of the group under study are unaware of the structural models in terms of which the anthropologist interprets constellations and social relations».

Miller, Galanter and Pribram (1960) discussed *planning* performance in terms of a hierarchy of *tote* units. Such a unit consists of two parts: a *test* to see if some situation matches an internally generated goal and an *operation* that is intended to reduce any differences between the external situation and some internally stated goal. The goal may derive from a model or hypothesis about what will be perceived or what would constitute a satisfactory state of affairs. The operations can either revise the goal in the light of new evidence received or they can lead to actions that change the organism's internal and/or external environment. The test and its associated operations are actively linked in a feedback loop to permit iterated adjustments until the goal is reached. Such a *tote* unit is in some ways a forerunner to our own *planning machine* and is represented in the same way (see fig. 7.4).

The *internal* structure of the machine - which is the principal purpose of *our* work - is not explored in any detail in their works.

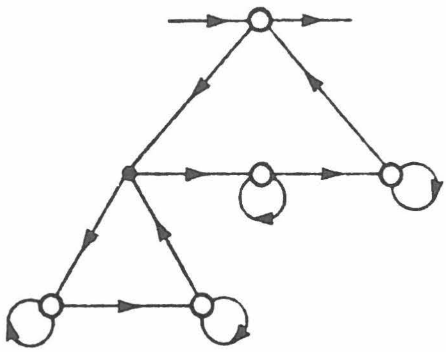


Figure 7.4 A possible communication links of tote units

Finally, we note two points about studies of competence and performance as understood in Chomsky's work (March, 1979):

«(1) studies of design [planning] performance need to be based on assumptions about underlying competence - the latter having priority over the former, and...

(2) theories of competence cannot be extracted from what designers [planners] appear to know or think they know about design [planning]».

References

Ackoff R.L. (1974) Beyond Problem Solving, *General Systems, 19*, 237-239.

Aleksandrov P.S. (1956) *Combinatorial Topology*, Graylock Press, Rochester, N.Y..

Arbib M.A., Manes E.G. (1975) *Arrows, Structures and Functions: the Categorical Imperative*, Academic Press, New York.

Atkin R.H. (1974) *Mathematical Structure in Human Affairs*, Heinemann Educational Books, London.

Atkin R.H. (1977) *Combinatorial Connectivities in Social Systems*, Birkhauser Verlag, Basel and Stuttgart.

Batty M. (1974) A Theory of Markovian Design Machines, *Environment and Planning B, 1*, 125-146.

Batty M. (1975) Design as Collective Action, *Environment and Planning B, 2*, 151-176.

Batty S.E. (1977) Game-Theoretic Approaches to Urban Planning and Design, *Environment and Planning B, 4*, 211-240.

Berry B. (1973) *The Human Consequences of Urbanization*, MacMillan, London.

Bobrow L.S., Arbib M.A. (1974) *Discrete Mathematics: Applied Algebras for Computer and Information Science*, Saunders, Boston (chapter on *Machines in a Category*, 664-707).

Boguslaw R. (1965) *The New Utopians: a Study of Systems Design and Social Change*, Prentice Hall, Englewood Cliffs, NJ.

Burks A.W. (1977) *Chance, Cause, Reason*, University of Chicago Press, Chicago.
Centre de Mathematique Sociale Ecole des Hautes Etudes en Sciences Sociales (1976) *Combinatorics, Graphs and Algebra*, Mouton, The Hague.
Chaitin G. (1975) Randomness and Mathematical Proof, *Scientific American, 232*, 47-54.
Chomsky N. (1963) Formal Properties of Grammar in Luce R.D., Bush, R.R., Galanter E. (eds.) *Handbook of Mathematical Psychology*, Wiley, New York, 323-418.
Chomsky N. (1965) *Aspects of the Theory of Syntax*, MIT Press, Cambridge, Mass.
Coleman J.S. (1964) *Introduction to Mathematical Sociology*, Collier-Macmillan, London.
Cortés F., Przeworski A., Sprague J. (1974) *Systems Analysis for Social Scientists*, Wiley-Interscience, New York.
Couclelis H. (1977) Urban Development Models, PhD Thesis, University of Cambridge.
Craik K.J.W. (1943) *The Nature of Explanation*, Cambridge University Press, Cambridge.
Dunn E.S. (1974) *Social Information Processing and Statistical Systems - Change and Reform*, Wiley-Interscience, New York.
Gel'fand I.M., Tsetlin M.L. (1971) Mathematical Modelling of Mechanisms of the Central Nervous System in Gel'fand I.M. (ed.) *Models of the Structural-Functional Organization of Certain Biological Systems*, MIT Press, Cambridge, Mass., 1-22.
Graver J.E., Watkins M.E. (1977) *Combinatorics with Emphasis on the Theory of Graphs*, Springer-Verlag, New York.
Gurvitch G. (1971) *The Social Frameworks of Knowledge*, Basil Blackwell, Oxford.
Hall A.D., Fagen R.E. (1956) Definition of System, *General Systems, 1*, 18-28.
Harary F., March L., Robinson R.W. (1978) On Enumerating Certain Design Problems in Terms of Bicoloured Graphs with no Isolates, *Environment and Planning B, 5*, 31-43.
Harary F., March L., Robinson R.W. (forthcoming) Bicoloured Graphs: Design Generation and the Value of Theory.
Harary F., Norman R.Z., Cartwright D. (1965) *Structural Models: an Introduction to the Theory of Directed Graphs*, Wiley, New York.
Harris B. (1978) A Note on Planning Theory, *Environment and Planning A, 10*, 221-224.
Howson A.G. (1972) *A Handbook of Terms Used in Algebras and Analysis*, Cambridge University Press, Cambridge.
Klir G.J. (1969) *An Approach to General Systems Theory*, Van Nostrand Reinhold, New York.
Knuth D. (1973) Computer Science and Mathematics, *American Scientist, 61*, 709-719.
Kron G. (1963) *Diakoptics: the Piecemeal Solution of Large-Scale Systems*, MacDonald, London.
Kuratowski K. (1972) *Introduction to Set Theory and Topology*, Polish Scientific Publishers, Warsaw.
Lancaster K.J. (1966) A New Approach to Consumer Theory, *Journal of Political Economy, 74*, 132-157.
Lancaster K.J. (1971) *Consumer Demand*, Columbia University Press, New York.
Lefschetz S. (1975) *Applications of Algebraic Topology*, Springer-Verlag, New York.
Leonard H.S. (1957) *Principles of Right Reason*, Holt, New York.
Löfgren L. (1972) Relative Explanations of Systems in Klir G.J. (ed.) *Trends in General Systems Theory*, Wiley-Interscience, New York, 340-407.
MacLane S. (1972) *Categories for the Working Mathematician*, Springer-Verlag, New York.
Manes E.G. (ed.) (1973) *Category Theory Applied to Computation and Control*, Springer-Verlag, New York.
March L. (1975) The Logic of Design and the Question of Value in March L. (ed.) *The Architecture of Form*, Cambridge University Press, Cambridge, 41-73.
March L. (1979) Editorial, *Environment and Planning B, 6*, 1-2.
March L., Batty M. (1975) Generalized Measures of Information, Bayes' Likelihood Ratio and Jaynes' Formalism, *Environment and Planning B, 2*, 99-105.
March L., Ho Y.-S. (1976) Reviews on: Atkin R.H. *Mathematical Structure in Human Affairs*; Cortés F., Przeworski A., Sprague J. *Systems Analysis for Social Scientists*; Hartigan J.A. *Clustering Algorithms*, *Environment and Planning B, 3*, 135-141.

Miller G.A., Galanter E., Pribram K. (1960) *Plans and the Structure of Behaviour*, Holt, New York.

Piaget J. (1971) *Structuralism*, Routledge and Kegan, London.

Pontryagin L.S. (1952) *Foundations of Combinatorial Topology*, Graylock Press, Rochester, N.Y..

Rosen R. (1978) *Fundamentals of Measurement and Representation of Natural Systems*, North-Holland, New York.

Schreider Ju.A. (1975) *Equality, Resemblance and Order*, MIR Publisher, Moscow.

Scott A.J. (1971) *Combinatorial Programming, Spatial Analysis and Planning*, Methuen, London.

Simon H.A. (1965) The Logic of Rational Decision, *British Journal for the Philosophy of Science, 16*, 169-186.

Simon H.A. (1967) The Logic of Heuristic Decision Making in Rescher N. (ed.) *The Logic of Decision and Action*, University of Pittsburgh Press, Pittsburgh, 1-20.

Simon H.A. (1969) *The Sciences of the Artificial*, MIT Press, Cambridge, Mass.

Simon H.A. (1977) *Models of Discovery*, D. Reidel, Dordrecht.

Simon H.A., Newall A. (1972) *Human Problem Solving*, Prentice Hall, Englewood Cliffs, N.J..

Solomonoff R. (1964) A Formal Theory of Inductive Inference, *Information and Control, 7*, 25-36.

Stedman Jones G.S. (1973) History: the Poverty of Empiricism, in Blackburn R. (ed.) *Ideology in Social Science*, Fontana/Collins, London, 96-115.

Stiny G., Gips J. (1978) *Algorithm Aesthetics: Computer Models for Criticism and Design in the Arts*, University of California Press, Berkeley.

Stiny G., March L. (1981) Design Machine, *Environment and Planning B, 8*, 245-255.

Stone R. (1967) *Mathematics in the Social Science and Other Essays*, Chapman Hall, London.

Tonti E. (1975) On the Formal Structure of Physical Theories, Istituto di Matematica, Politecnico di Milano, Milano.

Tribus M. (1969) *Rational Descriptions, Decisions and Designs*, Pergamon Press, London.

Wright G.H. von (1962) *The Logic of Preference*, Edinburgh University Press, Edinburgh.

Wright G.H. von (1963) *Norm and Action*, Routledge and Kegan, London.

Wright G.H. von (1967) The Logic of Action - a Sketch in Rescher N. (ed.) *The Logic of Decision and Action*, University of Pittsburgh Press, Pittsburgh.

Zeeman E.C., Buneman O.P. (1968) Tolerance Spaces and the Brain in Waddington C.H. *Towards a Theoretical Biology 1 Prolegomena*, Edinburgh University Press, Edinburgh.

Appendix

Categorical algebra represents a fundamental approach to the study of such abstract structures as groups, rings, fields, topological spaces, etc. The emphasis in this approach lies not so much on individual groups or rings, say, but on the connections between similar structures and the functions which relate them. The notion of a category was first introduced in a topological setting by S. MacLane and S. Eilenberg in 1945. Since then the concept has been extended to help unify other aspects of mathematics.

A **category** C is a class of objects, A, B, C, ..., together with two functions:

(a) a function which assigns to each pair of objects A, B $\in C$ a set C(A,B) (or hom (A,B)) called the *set of morphisms* with domain A and codomain B;

(b) a function assigning to each triple of objects A, B, C $\in C$ a law of composition

$$C(A, B) \times C(B, C) \to C(A, C) ,$$

(if f $\in C$(A, B) and g $\in C$(B, C), then the composite of f and g written gf, is a morphism with domain A and codomain C).

The functions and morphisms so defined must, moreover, satisfy the axioms:

I. $C(A_1, B_1)$ and $C(A_2, B_2)$ are disjoint unless $A_1 = A_2$ and $B_1 = B_2$;

II. (Associativity) Given f $\in C$(A, B), g $\in C$(B, C) and h $\in C$(C, D), then (hg) f = h (gf);

III. (Identity) To each object A in C there is a morphism $1_A : A \to A$ such that, for all f $\in C$(A, B), g $\in C$(C, A), we have

$$f 1_A = f , \quad 1_A g = g .$$

In the particular case when, to each of the objects A, B, C, ..., there corresponds a set U(A), U(B), ... known as the 'underlying set' of A, B, ..., and the elements of C(A, B), etc. are functions from the set U(A) to the set U(B), C is called a **concrete category**.

If the class of objects A, B, C, ... forms a set, then C is called a **small category**.

A morphism f $\in C$(A, B) is said to be **invertible** (a **unit** or an **isomorphism**) in C if there exists g $\in C$(B, A) such that

$$gf = 1_A , \quad fg = 1_B .$$

If f is invertible, we write $g = f^{-1}$ and say that A and B are **equivalent** in C. A category in which every morphism is invertible is known as a **groupoid** (but see note on p. 25).

A morphism m $\in C$(B, C) is said to be **monic** in C if, for all f, f′ $\in C$(A, B) and all A in C, mf = mf′ implies f = f′.

A morphism e $\in C$(A, B) is **epic** in C if ge = g′e for two morphisms g, g′ $\in C$(B, C) always implies g = g′.

An object I in C is said to be an **initial object** in C if for all X in C there is exactly one morphism I $\to$ X, i.e. if C(I, X) is a singleton (p. 9) for all X in C. Similarly an object T in C is called a **terminal object** if C(X, T) is a singleton for all X in C. An object which is both initial and terminal is known as a **zero object**.

If C possesses a zero object P, then, for all A, B in C, C(A, B) contains the morphism gf where $\{f\} = C(A, P)$ and $\{g\} = C(P, B)$, gf is independent of the choice of zero object and is known as the **zero morphism**. It is denoted by $o_{AB} : A \to B$.

Given two categories, C and D, we define a **functor** $F : C \to D$ to be a pair of functions, namely,

(a) the *object function* that assigns to each A in C an object F(A) in D; and

(b) the *mapping function* that assigns to each morphism $f : A \to B$ of C a morphism $F(f) : F(A) \to F(B)$ of D, which satisfy

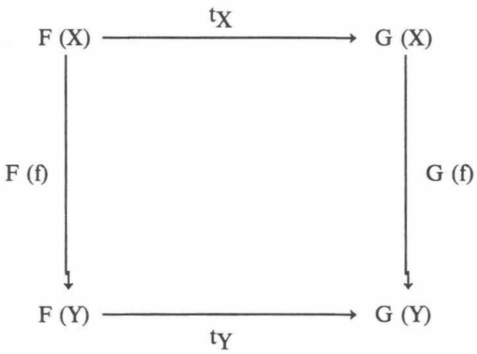
I. $F(1_A) = 1_{F(A)}$ for each A in C,

II. $F(gf) = F(g)\,F(f)$ for each composite gf defined in C.

Given two functors $F, G : C \to D$ then a **natural transformation**: $F \to G$ is a function which assigns to each object $X \in C$ a morphism $t_X : F(X) \to G(X)$ in D in such a way that every morphism $f : X \to Y$ of C yields a commutative diagram, i.e. that

$$G(f)\,t_X = t_Y\,F(f) .$$

If t_X is invertible for each X, then t is called a **natural isomorphism** or **natural equivalence**.



Résumé. Cet essai passe en revue les différents types de théories et les formes du processus de planification. Il présente ensuite un cadre du processus de formation d'un plan («machine de planification») basé sur des conceptions dérivées de la théorie de l'information. L'importance de la théorie des catégories par rapport à la notion de «machine de planification» est mise en évidence et l'on discute ensuite l'application de cette conception aux différentes approches de planification et aux différents contexts institutionnels.

Riassunto. Questo articolo passa in rassegna diversi tipi di teorie e di forme del processo di pianificazione; presenta, inoltre, un quadro del processo di formazione di un piano («macchina del piano»), basato sui concetti derivati dalla teoria dell'informazione. Viene messa in luce, con opportuno dettaglio, l'importanza della teoria delle categorie per la concezione di «macchina del piano», e viene discussa l'applicazione della suddetta concezione nei diversi approcci di pianificazione e nei diversi contesti costituzionali.

Environment and Planning A

1982, volume 14, number 8

Environment and Planning A

1982, volume 14, number 9

Special issue of papers from Germany and Austria

Environment and Planning A

1982, volume 14, number 10

Environment and Planning A

1982, volume 14, number 11

Environment and Planning A

1982, volume 14, number 12

Environment and Planning B

1982, volume 9, number 2

INFORMATION FOR THE AUTHORS

1. Papers can be submitted in one of the following languages; Italian, French, English. As a general rule, papers should not be longer than 40 typewritten pages, including References (with ca. 30 lines per page and 2 cm. margins). Each paper must be proceded by a summary on the aims and the results achieved, of not more than 200 words, also providing a maximum of 5 key words.
   The name of the author (or of each author) must be followed by the name and address of the institution or organization he belongs to.
   At the beginning of each paper, introductory remarks must illustrate the problem being dealt with and the author's main contributions.
   Manuscripts must be sent, in three copies, to the Director.
   C.S. BERTUGLIA
   Istituto di Scienza dei Sistemi Architettonici e Territoriali
   Facoltà di Architettura - Viale P.A. Mattioli, 39 - 10126 TORINO, ITALIA

2. Each paper will be examined by two members of the Scientific Committee Actual publication will take place provided the author has accepted any suggestions by the Director or the two Scientific Committee members, transmitted in the author by the Director.
   Manuscripts will not be returned.

3. All figures (including tables, diagrams, etc.) must be numbered consecutively in the order in which they are cited in the text.
   Tables and diagrams must be in a form which can be readily reproduced, of 12,5×19 cm. maximum size.

4. Mathemathical formulae must be closely spaced, as in the example below:
   $F_{ij}=[\Sigma_j N_j D_j exp(-\beta c_{ij})]^{-1} N_j O_j D_j exp(-\beta c_{ij})$.
   Letters denoting matrices and vectors must be underlined with a double line.
   When necessary, mathematical formulae must be numbered consecutively with arabic figures in round brackets.

5. Bibliographical references included in the text must give the author's last name (or author's) followed by the year of publication, as in the following examples: (Shannon and Weaver, 1949) or Shannon and Weaver (1949).

   When referring to more than one publication of the same year by one author, they should be identified by letters of the alphabet, as follows: (Hall, 1976a) or Hall (1976a).

   The bibliography must be at the end of the paper, giving, for books: last names and initials of all the author's names, the year of publication, complete title, publisher and city, as in the example below:
   March L. (1967) A Boolean description of a class of built forms, in L. March (ed.) *The Architecture of Forms*, Cambridge University Press, London, 41-73.

   For journals: last names and initials of all the author's names, the year of publication, complete journal title, volume number and pages, as in the following example:
   Lakshmanan T.R. and Hansen W.G. (1965) A retail market potential model. *Journal of the American Institute of Planners*, 31, 134-143.