SPATIAL DYNAMICS AND METROPOLITAN CHANGE

Editor:
Börje Johansson
Centre for Regional Science Research (CERUM), University of Umeå,
S-90187 Umeå, Sweden

RR-86-9
November 1986

Reprinted from *Regional Science and Urban Economics*, **16**, 1–160 (1986).

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
Laxenburg, Austria

Reprinted with permission from *Regional Science and Urban Economics*, **16**, 1–160 (1986).

Printed by Novographic, Vienna, Austria

FOREWORD

After several decades of general metropolitan growth, the end of the 1960s marked the emergence of a more dispersed situation, with a significant decline in metropolitan regions dominated by mature and obsolete industries and a simultaneous vitalization of creative agglomerations, functioning as centers for knowledge creation, product development, negotiations, and similar activities. These changes, which reflect basic technological transitions in the world economy, are causing profound structural adjustments within the regions, as well as in their external trade and contact patterns.

These structural adjustments have been studied and analyzed in the Project "Dynamics of Metropolitan Processes and Policies", initiated within the Regional Issues Group at the International Institute for Applied Systems Analysis (IIASA). The Project was organized as a comparative and collaborative effort of research groups in about 20 metropolitan regions. The approach of the still active network is two-pronged, including:

(1) Empirical comparisons of change patterns.
(2) Development of theories, models, and methods suited for the analyses of metropolitan dynamics.

This *Research Report* consists of contributions belonging to the second category, earlier versions of which have appeared as IIASA Working or Collaborative Papers. Contributions belonging to the first category are collected in *Dynamics in Metropolitan Processes and Policies* (RR-86-8), also edited by Börge Johansson.

BORIS SEGERSTAHL
Deputy Director
International Institute for Applied Systems Analysis

CONTENTS

Regional Science and Urban Economics 16 (1986) 1–6. North-Holland

SPATIAL DYNAMICS AND METROPOLITAN CHANGE

Introduction

Börje JOHANSSON

CERUM, University of Umeå, 901 87 Umeå, Sweden

This special issue presents contributions to a collaborative effort to analyze 'the dynamics of metropolitan processes and policies'. That effort was initiated at IIASA (International Institute for Applied Systems Analysis) and is aiming at (i) empirical comparisons of change processes in a set of metropolitan regions located in different parts of the industrialized countries, in combination with (ii) theoretical development of models and methods suited for dynamic analyses of metropolitan processes.[1]

This volume contains four papers which focus on industrial change and economic restructuring; two papers deal with population relocation and migration processes; one paper contains a study of economic cycles in space and one paper treats the assessment of urban investment and urban renewal projects.

Metropolitan change

Metropolitan regions constitute the nodes of the interregional network of nations and of the international system; they also encompass a significant share of the economic activities in the world economy. Moreover, they often function as centres for business and governmental decisions, negotiations, knowledge creation and other face-to-face activities.

Over time a metropolitan region is forced to adapt its internal structure in response to external economic and demographic changes as well as long-term technological development. This aspect of metropolitan change is outlined in the paper by Lakshmanan and Chatterjee. They describe the advent of the mature metropolis in a historical context of technological change and structural adjustments in urban infrastructure. Thereby they also provide a problem-oriented background of the entire metropolitan study.

[1]The objectives and approaches in the collaborative study are summarized in Johannson (1985) and Snickars (1985).

0166–0462/86/$3.50

Adjustments in the form of urban renewal and other investment projects are investigated in the paper by Quigley which examines methods to assess such projects. The internal processes of change include complex dynamics of spatial relocation, household formation and incongruencies between supply and demand of capacities in the transportation, housing, service and production systems. The location and spatial distribution of urban population is analyzed in the paper by Leonardi and Casti, in which an explorative extension of a static to a dynamic framework is undertaken. In this analysis the urban region is viewed as a public good; the agglomerative force is derived from the propensity of people to interact, and this force is counteracted by the limited amount of urban land that is available.

Studies of urban development during recent decades indicate that technology and knowledge intensive industries frequently locate in regions with a rich variety of education, research and cultural opportunities. In the paper by Anderstig and Hårsman this phenomenon is examined both in a multiregional and intraregional perspective. Referring to a product cycle theory, metropolitan regions are studied as birth places for new technologies, and this is related to the competence profile of the labour force and the intensity of knowledge oriented occupations in each region.

The Anderstig–Hårsman paper also analyzes the changing patterns of intra-urban location of economic activities. A similar problem is studied by Hayashi and Isobe but from a different perspective. In this paper, the location and relocation of firms is modelled with the help of a nested logit framework which takes into accout both a firm's attributes and the characteristics of alternative locations including properties of the transportation system.

The location of different types of production activities is further emphasized in the paper by Camagni, Diappi and Leonardi in a simulation model of urban growth and decline in a spatial system with many urban centres. The process of change is influenced by location benefits associated with an urban hierarchy and the appearance of innovations.

A multiregional setting is also found in the Haag–Weidlich model of migration processes; this model provides a theoretical framework for analyzing interregional population flows with a stochastic specification.

The paper by Puu contains a study of economic interaction in continuous space. In this case cycles over time and space are generated by a multiplier–accelerator principle for the economic process, referring to an economy with interregional trade and non-linear investment responses.

Spatial dynamics

A common element of urban change processes is the inertia in the interprocess adjustment mechanisms. As housing is constructed in peripheral rings to accommodate an increased population, the pressure on the land in the down-town business district may accelerate. The household and work-place relocation process brings about tensions in e.g., the land market and transportation system. The tensions and the associated signals (such as prices) give rise to adjustment of different time scales.

A central issue in the metropolitan study has been the differentials as regards the speed of adjustments of various subsystems (or variables). As described in fig. 1, the relative speed of change is an ambiguous concept. The figure describes an investment cycle of the type modelled in the paper by Puu. Obviously, in the depicted process which converges towards (Y^0, I^0), I changes much faster than Y on the curve segment A–B, while the opposite is true for the segment B–C. Therefore, other ways of characterizing the speed of change have been considered. One is represented by the following classification:

- A system following a steady path of balanced growth, in which the speeds of change of its various components are balanced against each other. Such a change process develops along a trajectory that may be thought of as an equilibrium path.
- A system of the above type may suddenly be affected by strong exogenous changes. This may bring about repercussions in the form of a faster speed of change in some parts of the system, e.g., adjustments towards a new steady path.

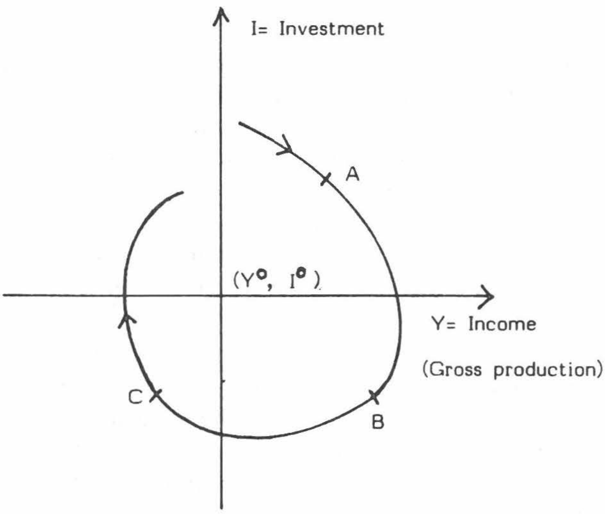


Fig. 1. Illustration of relative speed of change.

– Catastrophic shifts in the speed of change based on bifurcations and singularities. Fig. 2 illustrates a special case in which such shifts occur repetitively (cyclically), with long duration of the slow phases [see e.g., Casti (1985)].

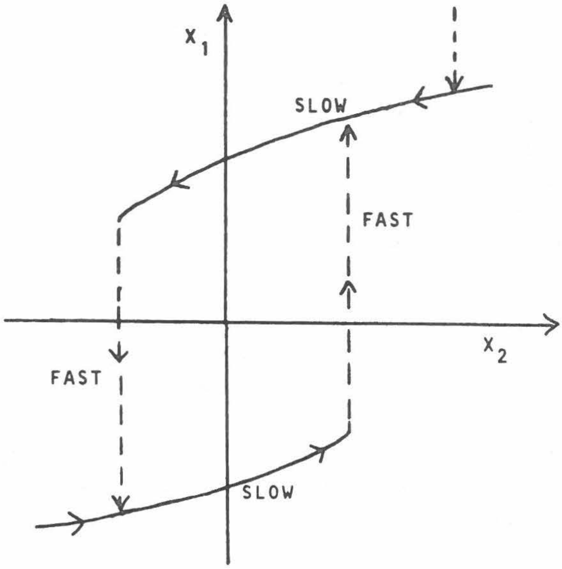


Fig. 2. A cycle of fast and slow time scales.

In relation to the change path in fig. 1 we may also mention the possibility of describing that process by decomposing time into two variables – one fast time representing the cyclic motion and one slow time for the trend factor associated with the asymptotic approach to a limit cycle or a stable equilibrium point (cf. the paper by Puu).

A second classification of time scales can be based on for how long time a phenomenon or an object remains in the same spatial position, i.e., the locational duration. Snickars (1985) suggests a division into three time-layers. Long-term change and slow adjustments of land-use patterns correspond to the development tempo of the fixed urban capital built on land. The medium-term layer refers to the locational dynamics of the economic and demographic processes taking place within the built-up stocks and urban structure. The short-term time scale corresponds to flows in the regional setting that bring about the concrete interaction between activities with given location. Related to this time scale are transportation of commodities, workplace commuting etc. When these processes are observed on the short-term scale they display an oscillatory behaviour. On the medium-term scale the same processes often are represented by smooth equilibrium flows.

When models of 'slow' or 'medium speed' changes are constructed, processes observed on a faster time scale usually are embedded in the processes of the less fast time-scale. Processes operating on a slower time-scale, on the other hand, usually appear in the form of parameters for processes modelled on a faster time-scale. When such parameters gradually change, they may cause bifurcations in models with non-linear behaviour [see e.g., Varaiya and Wiseman (1984)].

The locational duration of the long- and medium-term layers may be described as in fig. 3. Here the paths A and B may represent the development of demand for a certain capacity in a given location. The stair-case curves, A^* and B^*, represent the slow and discontinuous adjustments in response to signals of tension. The step-like character of the adjustments may be related to delays and rigidities of the decision system, sunk costs effects as well as durability and indivisibilities of spatial structures. We may finally observe that spatial aggregation of different locations into large zones and regions will make aggregate versions of the A^* and B^* curves more smooth. In this sense we may observe a certain similarity between spatial aggregation and 'time aggregation', where the latter represents a shift from a faster to a slower time-layer.

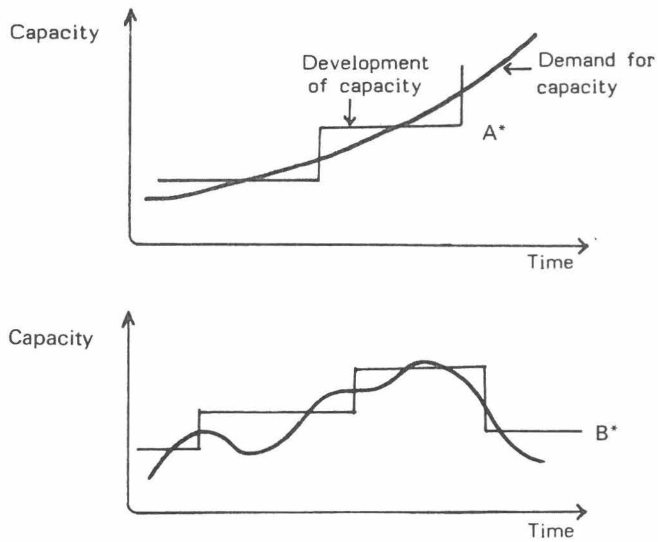


Fig. 3. Slow adjustments of capacity in a given location.

The contributions presented in this volume cover only some of the issues and theoretical suggestions outlined in this introduction. In fact, these introductory observations rather represent research ambitions, to some extent even very long-term ambitions. However, the selection of issues and the problem formulations in this volume have been strongly influenced by the above types of dynamic considerations. It is a challenging effort to further stimulate a development with a combination of problem formulation, model construction and design of methods for empirical observations along these lines.

References

Casti, J., 1985, Simple models, catastrophes and cycles, Research report RR-85-2 (IIASA, Laxenburg, Austria) (see also Kybernetes 13, 1984).

Johansson, B., 1985, Dynamics of metropolitan processes and policies, Scandinavian Housing and Planning Research 2, no. 3–4.

Snickars, F., 1985, Beyond comparative analyses of metropolitan dynamics, Scandinavian Housing and Planning Research 2, no. 3–4.

Varaiya, P. and M. Wiseman, 1984, Bifurcation models of urban development, in: A.E. Andersson, W. Isard and T. Puu eds., Regional and industrial development theories, models and empirical evidence (North-Holland, Amsterdam) 61–88.

Regional Science and Urban Economics 16 (1986) 7–30. North-Holland

TECHNICAL CHANGE AND METROPOLITAN ADJUSTMENTS

Some Policy and Analytical Implications*

T.R. LAKSHMANAN and Lata CHATTERJEE

Boston University, Boston, MA 02215, USA

Received March 1985, final version received August 1985

The objective of this paper is to highlight important elements of models of metropolitan evolution in developed countries and thus outline the emerging agenda of metropolitan modeling. The advent of the mature metropolis is analyzed in a historical context of technical change, structural adjustments, and rigidities in change management. Four classes of issues are addressed: (i) dynamic adjustment processes, (ii) distributional issues, (iii) links of the urban system to the nation, and (iv) the modeler–decision maker system.

1. Introduction

A remarkable restructuring of metropolitan space is underway in the industrialized countries. After two decades of rapid overall growth and peripheral expansion, the metropolitan areas are undergoing a transformation usually described as the onset of maturity. The characteristics of maturity include population declines, accelerating industrial evolution, spatial shifts and the diffusion of the functioning urban areas beyond the defined boundaries of metropolitan areas.

On a little reflection, one would note that these metropolitan changes are accompanied by an emerging broad restructuring of geographic space at national and global levels. Gentrification and extra metropolitan growth, deindustrialization and regional decline, the industrialization of the Third World and a recent international division of labor, world wide dislocations and neoprotectionism – these are not isolated developments but symptoms of a world wide transformation. The geographical restructuring reflects deep-seated changes in the broader socioeconomic context – just as urbanization

*Earlier versions of this paper were presented at the IIASA Workshop in Rotterdam in June 1984, at the Universities of Wales and Sheffield, an EEC Workshop on 'Technology Change and Employment', and at the Academy of Economics, Poznan, in early 1985. The authors are grateful to many individuals for their comments – particularly Chang-i Hua, Ake Andersson, Michael Batty, Philip Cooke, Patrick O'Farrell, Ian Messer, Börje Johansson, Bohdon Gruchman and an anonymous referee.

and the modern city developed in response to major structural changes in the broader society in an earlier period.

The modern metropolis represents an adjustment to and a facet of the industrial revolution of the last century. The emergence of the factory system with its associated scale and external economies, in the context of the dominant technology of the age (steam engine and intercity railroads) led to spatial concentration of activities. In the last half century, a variety of developments in physical and institutional technology – the motor car, the truck, the highspeed road, information technology, and cheap mortgage credit, etc. – greatly weakened the hold of urban areas on economic activities and led to the patial spread in urban space. Demographic trends and shifts in values appear to accentuate further the recent trends toward metropolitan maturity.

In a similar vein, future metropolitan outcomes can be viewed as reflecting the interplay of a variety of technological economic and social changes and adjustments to such changes on the part of households, firms and institutions. Such adjustments will be powerfully influenced by the forms and types of public intervention.

It is our view that metropolitan models, if they are to retain serious intellectual and policy interest, must begin to address such themes relevant to the context and scope of metropolitan evolution. Processes of change, adjustment, and facilitating public interventions – should be on the agenda of metropolitan modeling. The objective of this paper is to highlight important elements of models of metropolitan evolution in developed countries and thus outline the emerging agenda of metropolitan modeling.

Since we argue that the basic force is technical change and adjustments to it, in terms of the organization and location of urban activities, cannot be separated from the broader economic and political setting within which they take place, we begin section 3 by outlining two complementary interpretations of the processes of long-term change in industrialized countries. We proceed in section 4 to a discussion of successful examples of adjustments to technical change as well as the significant rigidities to structural adjustment in affluent countries. Finally, we offer some speculations on the likely effect of these processes of technical change and adjustment on the functional and spatial organization of urban activities.

We turn in section 5 to the role of the public sector in facilitating these adjustments. Since the areas and forms of public intervention vary with the conditions and requirements of the times, our interpretation of the requirements of adjustment to change helps us identify some likely 'postures' and forms of public intervention that would reduce costs of adjustment.

We proceed in section 6 to translate the previous discussion on the context and scope of metropolitan evolution into a few questions that ought to be on the agenda of urban models, if the latter are to be of interest to urban theory and policy. We recognize four classes of issues not normally addressed in urban models – dynamic adjustment processes, distributional issues, links of the urban system to the nation, and the modeler–decision maker system.

Finally, we note that the thrust of our paper has been to highlight the context and scope of metropolitan change in order to identify the sort of questions that need to be embodied in mathematical models of the urban system. In this process we have adopted the ambitious and high-risk strategy of ranging over broad areas of technical change, socioeconomic adjustments and the role of the public sector – all of which are grist to the modeler's mill. In this process, some readers will find us guilty (infrequently we trust) of gross simplifications and of some highly contestable statements. In the usual choice in these occasions between being precisely wrong and vaguely right, at least our intentions must be clear.

2. The advent of the mature metropolis

The modern metropolis is the culmination of the industrial evolution of the 19th century that was largely directed to production, distribution and consumption of goods. The underlying features of this industrial development were the factory system, and its associated internal economies of scale, and urbanization economies that derived from the complementarities of different production activities. The steam engine and the intercity railroad led initially (in the context of the high cost of intracity freight transport) to spatial concentration of economic activities in cities. However, in the last 50 years, a variety of technological and institutional developments – the motor car and truck, high speed expressways, cheap mortgage credit, technological evolution towards a greater role for services, and technical developments in information storage, transmission and retrieval – have greatly attenuated these urban linkages of industries. Other developments – the expansion of the service sector (given the high income elasticity of services), the declining relative need for goods handling (given the increasing technical efficiency of materials used in production) and the roles of the motor truck and passenger car in reducing costs of intra-urban freight and commuter transport, etc. – further eroded the attractions of the central locations on economic activities. The overall effects of all these developments is the enormous weakening of the hold that central cities and even the metropolitan areas have on economic activities and the increasing tendency of goods production and service activities to avoid the high costs of congestion and urban public services in urban areas and to move into areas beyond the boundaries of metropolitan areas – leading to what is described as metropolitan stagnation or 'maturity'.

The demographic trends in the same period appear to reinforce the effects of the economic evolution. The major cause of population decline in metropolitan areas is the sharp drop in the birth rate and the consequent fall in natural increase. 'When small rates of population increase are coupled with continuing differential exchanges among metropolises, it is a mathematical certainty that many metropolitan areas will experience population loss' [Alonso (1978)]. These population losses partly reflect the spread of population beyond the boundaries of Standard Metropolitan Statistical Areas (SMSAs) into the countryside, a process that has been going on for a long time. Further, these numerical changes seem to be accompanied by shifts in demographic characteristics and social values [Leven (1978)].

While the characteristics of maturity vary from country to country, their metropolitan areas share the following attributes:

(1) population declines and spatial shifts of population,
(2) declining industrial jobs,
(3) physical decay in cities and growing bottlenecks in suburbs,
(4) diffusion of functioning urban areas beyond the defined boundaries.

Metropolitan areas in industrialized economies have been experiencing a decentralization trend for some decades. Till about a decade ago, the suburban sections of SMSAs have been growing much faster than the central cities. In the last decade or more, central cities have lost population and even entire metropolitan areas have stabilized or begun to decline in population while ex-urban locations experience population growth. This decentralization is well advanced in the U.S.A. where macro regional shifts in population from the 'frost belt' to the sunbelt accentuate these trends in the North East. In the decade of the seventies, central cities gained in population only by 1.3% while suburbs registered a 5.8% increase to yield an overall SMSA population gain of 3.7%. However, central cities lost population in the older industrial areas of North Central (10.1%) and North East (10.7%). (See table 1.) The loss of population in the central cities was so severe, that in spite of sizeable sururban population growth, the SMSAs in the North East lost population (2%). By contrast, the SMSAs in the growing regions of the sunbelt of the South and West increased in population (as did many of their central cities as well).

Table 1
Population changes in the U.S.A., 1970–1980 (% change in SMSAs).

Region	SMSAs	Central cities	Suburbs	Non-SMSA
U.S.A.	3.7	1.3	5.8	15.1
North East	−1.9	−10.7	4.1	13.5
North Central	2.7	−10.1	11.2	7.5
South	21.5	14.7	26.6	17.2
West	22.6	14.0	28.2	30.6

Source: U.S. Census Bureau 1970, 1980.

In a similar fashion, the large metropolitan areas in the U.K., in the Randstad (Amsterdam, Rotterdam and The Hague) and West Germany have begun to lose population and decentralize further in the seventies. Swedish urban areas have a stable metropolitan area population but show a rapid loss of central city population. Indeed, this trend is evident in a number of OECD countries [Wolman (1983)].

What do these interconnected developments in demography, technology, economy and social values imply for the further evolution of the mature metropolis and its built environment? To begin to address this question, we need some understanding of technological, economic and social forces behind recent societal developments and some disciplined speculation of their likely evolution in the medium term – a task to which we turn next.

3. The emerging socio-economic context

Two competing formulations of the processes of long-term and progressive change in the world economy are relevant here. First is the intellectual curiosity in long-wave theory, the location of the last quarter of the twentieth century in such a formulation and the implications of that theory for designing adjustment processes. The second view is a structural formulation of economic development processes that provides some clues to the prospects for long-term growth and the types of adjustment processes necessary in the future. We proceed to a review of both these formulations.

3.1. The long-wave theory

A useful way to view the long-wave theory is to accept Van der Zwan's (1979) notion that there have been periodic major crises of adjustment, varying a little in their severity and timing between countries and followed by periods of expansion and prosperity that are perceived as far more severe than the usual downturns and upturns of business cycles. The first formulation of the long-wave theory is attributed to Kondratiev (1935) who, while analyzing long-term indicators, discovered several long waves of 50 to 60 years duration in the world economy. He suggested that inventions that were around would find application and initiate a major wave of economic expansion.

Schumpeter's (1939) contribution to this literature lies in his attribution of a central role to technical change in long-wave formation. Behind each Kondratiev cycle lay a set of major technical innovations, brought into the market by the ability of innovative entrepreneurs, who in turn attracted a 'swarm' of imitators and improvers to exploit the opportunities with a wave of new investment, generating boom conditions. The competitive processes set in motion by these swarming, imitative entrepreneurs eroded the profit margins (à la Marx). However, before a low level equilibrium trap is reached, the destabilizing effect of a new wave of innovations would start the process again. Schumpeter in this manner stressed the role of steam power in the first Kondratiev cycle (1818–1842), railroads in the second (1843–1897) and electric power and the automobile in the third (1898–1949). A neo-Schumpeterian would interpret the boom in the quarter century following 1948 as resulting from several technologies – electronics, synthetic materials, drugs, oil, petrochemicals and consumer durables.

Thus in the Schumpeterian framework, the causation runs from science via technology to the economy. The cumulative exploitation of these scientific and technological ideas is called the 'natural trajectory' [Freeman (1982)]. Different industries have different capacities to exploit these natural trajectories [Nelson and Winter (1977)]. In the Schumpeterian framework, it is disequilibrium and dynamic competition among entrepreneurs engaging in industrial innovation that are the basis of economic development. The focus is on the supply side, with autonomous investments rather than on demand processes driving development. Mensch (1979) also emphasized the supply side, claiming that basic inventions open up new investment opportunities that form the basis of whole new industries. He recognizes development periods in the last two centuries which have uniquely favored basic innovations. Mensch's data reveal that the innovation peaks precede the bottoms of Kondratiev cycles by two decades, suggesting that the beginning of a new boom is embedded in a previous downswing.

A second school of long-wave theorists emphasize the *demand* side factors. Chief among these are Schmookler (1966) who points out that once the major innovation is made, the role of the set of demand-led secondary inventions and innovations may be crucial for several decades; and the MIT Systems Dynamic Group focus on the role of demand for physical capital in the formation of long waves from the standpoint of large-scale economic shifts [Rothwell and Zegveld (1982)].

From the standpoint of large-scale economic shifts, what appears important is not the date of a particular basic technical innovation but a constellation of favorable circumstances – key technical inventions, the availability of large volumes of capital, the presence of entrepreneurs and a number of supporting managerial and institutional innovations. The importance of the managerial and institutional innovations (a subject we will return to in the next section) is evident from the fact that many applications of the steam engine required the reorganization of production on a factory basis, which represented a major excruciating social change at the time.

To understand the role of technical change on the downswing side, one must examine the nature rather than the rate of technical change in existing industries and the corresponding patterns of investment and demand growth. Utterback and Abernathy (1978) have suggested that the nature of innovations change over the duration of a growth cycle. As a new industry grows many new products are created that initially open up new markets and expand business. The product innovation phase is this 'expansionary' mode. As the product demand increases, investments in physical capital and in augmenting production efficiency are made. Eventually, the growth of demand slackens and a number of process innovations that reduce costs appear and the industry is in a 'rationalization' mode. With maturity, possibilities for new products decline and the industry begins to shed labor. Many 'smoke stack' industries (steel, petrochemicals, etc.) in the OECD countries appear to be in this advanced phase.

If this interpretation of the emerging context is correct, the source of new expansion in the next decade or two will lie in development of new markets in the developing world or more likely in the generation of new industries based on technologies currently in their infant or early stages of the development cycle. Potential examples of the latter are the biotechnologies, technologies related to energy and new resource frontiers such as the ocean floor. It has been suggested that these technologies could provide fresh investment opportunities, stimulate entrepreneurial activity and power the world economy into the next Kondratiev upswing [Rothwell and Zegveld (1982)].

The introduction of such technologies when the microelectronics – the key technology of our decade – is gaining momentum may be critical. New microelectronic control and steering devices provide flexibility and promote automation in small and medium-sized production units. Consequently, microelectronics will not only affect production but also private and public administration and in the service sector units. The potential for decentralized production and decentralized administration would increase [Friedrichs and Acheff (1982)]. Since administration planning, steering of production, and production can all have different locations, there are considerable implications for the future metropolitan built environment in such a scenario.

3.2. Structural model of development

The structural model of long-term progressive change advanced by A.G.B. Fisher (1935) and Colin Clark (1940) popularized the tripartite concept of economic structure (in terms of primary, secondary and tertiary sectors) and introduced the idea of structural change involving switching of resources from lower to higher productivity sectors as an indicator of and a reason for increasing productivity. The structural change that accompanies growth proceeds in stages: initially industry and services both increase as agriculture declines; in later stages agriculture reaches a minimum level, industry stabilizes and services continue to expand. The U.S.A. was the first 'service economy' – a transition from an industrial to a service economy in the 20th century deemed by Fuchs (1968) a revolution comparable in importance to the industrial revolution in England in the 18th century.

The reasons for the contemporary growth of the service sector in the Fisher–Clark model are

- Engel's Law – increasing incomes in the context of the higher income elasticity of services lead to greater demand for services,
- increasing division of labor in the production processes and the growth in producer services that are used as intermediate inputs to production, and
- the lower productivity of the service sector, which with the growth of the economy, shifts relatively more of the employment to the service sector.

While the service sector as a whole has been growing, different subsectors evidence different rates of growth. Following some of the functional notions of Browning and Singlemann (1978) and Gershuny and Miles (1983) we can recognize four service subsectors: distributive services (e.g., transport, communications, retail and wholesale), intermediate producer services (e.g., professional and technical services), final marketed services, and final (largely) non-marketed services (e.g., welfare, education, health, etc.). Of these, distributive services appear to be a stationary or slightly declining proportion of total employment, reflecting the differential growth rates of component categories. On the other hand, the intermediate producer services subsector is a dynamic sector, the demand for whose services is directly related to the increasing division of labor and the consequent rise in productivity in the manufacturing sector. Branches of the service sector such as banking and financial services that have grown rapidly recently evidence considerable productivity growth. By contrast, the long-term trends for the two types of final consumer services do not appear to be encouraging.

A major reason for this differential performance is the effect of relative prices of many services. In most countries, the rates of increase in the price of services are higher than the corresponding rates for all private consumption. In the U.K., for instance, between 1954 and 1974, the price of services rose 1.72 times as fast as that of durables [Gershuny (1978, p. 79)]. Longitudinally, the price effect acts against the income effect in a manner that lower than expected growth in service consumption takes place over time [Gershuny and Miles (1983)].

These price trends put pressure on the largely non-marketed final services subsector (e.g., education, welfare, health, etc.), that has been growing rapidly in the last two decades, characterized by both high labor-intensity and pressures for wage parity with workers in other sectors. As a consequence, there have been major pleas for the control of the scale of public expenditures – e.g., the California Proposition 13 and Massachusetts 2 1/2, both of which put a cap on state expenditures. Thus the long-term vitality of the service sector depends upon the ability to reduce costs in the final services subsectors. There is some evidence of this ability in innovative ways of provision of some marketed final services. (See next section.)

In summary, the preceding discussion of alternative interpretations of the processes of long-term change in the more affluent societies argues that the emerging world economy will likely reflect

- the cumulative effects of a cluster of new technologies, or
- the continuing structural evolution of the dominant service sector.

Indeed one may argue that these two formulations are not competitive but complementary explanations of the shape of things to come, since innovations in science, physical technology, organization and institutional development all interact in production and service delivery and thus determine the nature of the future world economy. Such a future will in all likelihood include the following characteristics: a plethora of new products; far ranging modifications brought about by the microelectronic information revolution in the organization of production and services; small and medium-sized production units increasing their potential for flexibility, control automation and outreach to markets; the greater spatial separation between various functions in workplaces – administration, planning, steering of production, production, and delivery of goods and services; the increasing restructuring of the service sector by bringing segments of economic activity into the 'informal economy' or the household; the consequent changes in household allocation of time and the organization of the household economy.

4. Adjustments to change and urban evolution

Major changes ensue from a combination of considerations. For example, steeply rising costs per unit of a major factor of production under conditions of growing demand lead often to technological change, or the availability of superior techniques provide incentives for technical change [Landes (1969)]. Whatever the origin, once underway, technological change is never smooth, since it activates on the one hand pressures to adjust societal structures and the resistance to structural adaptation on the other. Technical change calls for fundamental modifications. It involves the replacement of existing methods of production, service delivery and management. Consequently, it causes considerable damage to vested interests and often serious human dislocations. Thus rigidities to adaptations originate from the accumulation of institutions, rules and procedures which (instituted at an earlier period to achieve legitimate social objectives) are in the context of technical change a source of ineffectiveness in pursuing societal objectives [OECD (1981)]. Thus major changes are potentially two faceted; as disruptive tendencies that build up resistance to adaptation and as creative elements that increase the capacity for adjustments.

Insights into potential processes of adjustment to the future can be gained by an increased understanding of the dialectic between pressures for and rigidities to structural adaptation to technical change. We present below two examples of broad adjustments that have been made in the past to major changes.

4.1. Major adjustments: Two historical examples

In the latter part of the 18th century, British agriculture was exposed to the shock of cheap foreign competition (in a manner similar to that being faced by contemporary North Atlantic industrial production) which began to capture a growing share of the potential domestic and international market. Increases in farm productivity further reduced the demand for farm labor leading to structural unemployment. Subsequent developments in industry and transport (once the latter was mechanized by the introduction of the steam engine) offered enormous increases in productivity. The result of all of this was a redistribution of manpower and capital resources from the agricultural and rural to the industrial and urban sectors over the next several decades – the so called first industrial revolution.[1]

This massive restructuring became possible by a complex series of supporting technological, social, institutional, cultural and geographical developments – as in the move from the home or workshop to the factory, the substitution of a class society with its horizontal divisions for a vertical feudal structure, the occupational and geographical mobility of the workforce, etc.

The dynamic adjustments made to technical change in that era extended beyond those producing an improvement in physical capital. Equally important factors are [Hartwell (1971)]

- structural change (from agriculture, to industry and services),
- organizational change that occurred in all sectors resulting from better management, and
- investments in human capital (in the form of better nutrition, education, health, etc.) that made labor more productive.

To an important degree the above changes reflect the increasing role of the service sector (that had been growing in the U.K. and the U.S.A. from the beginning of their industrializations). The development of services played a key role in the first industrial revolution in three ways. These services provided

- *social overhead* capital which facilitated industrialization (e.g., transport,

[1]This transfer did not necessarily initially involve a reduction in absolute terms of output, manpower and capital engaged in agriculture.

communications, public administration, medicine, law and engineering professions),
- *intermediate services* which were necessitated by the increasing division of labor in production (e.g., banking, insurance, other financial services, real estate development, retail and wholesale trade, etc.),
- *cultural facilities* whose demand derived from increasing wealth (e.g., sports, entertainment, literature, journalism, music, etc.).

The key factor in the structural adjustment was the increasing productivity of services, especially of transport which enormously increased the market for products and the subsequent complementarity between increased industrial production and services. The increase in service productivity came from a combination of better physical capital (primarily transport), and better personnel (given the increased human capital investments and greater specialization of skills). Some historians of the industrial revolution suggest that in the U.K. and the U.S.A. the productivity in services was probably higher and was growing faster than in manufacturing till the latter part of the 19th century [Hartwell (1971)].[2] Only in the 20th century the growth of productivity in services in the advanced economies began to slow down and to rise more slowly than in manufacturing. There is some evidence, however, that suggests that productivity, while lower in the service sector than in manufacturing has experienced a recent acceleration [Levinson and Wheeler (1981)].[3]

This rationalization of some segments of the service sector through the reduction of labor costs is our second example of dynamic adjustment. This adjustment has arrived in the form of technical and organizational means of service provision to households. There is an increasing industrialization of the service sector. The cost reducing innovations involve the elements of subdivision of tasks, capital intensification, familiar economies of scale and the displacement of an important part of the service production outside the formal economy into the household. Manufactured consumer products

[2]The following table of structural change in the U.K. between 1750–1850 drawn from Hartwell (1971, p. 212), illustrates the higher productivity of services in early industrialization.

	Agriculture		Manufacturing		Services	
Year	% of national employment	Output	% of national employment	Output	% of national employment	Output
1750	45	45	30	25	25	30
1850	20	20	40	35	40	45

[3]Between 1870–1950 the ratio of rates of change of output/worker in service and manufacturing sector was 1/3. In 1950–1976, the same ratio had climbed to 1/2 [Levinson and Wheeler (1981)].

(autos, gasoline, T.V. sets, washing machines) are combined with intermediate service (e.g., repair services, T.V. programs), physical infrastructure (e.g., roads, broadcast networks, power transmission) and unpaid 'informal' labor (household labor) to produce personal services for transportation, and entertainment [Gershuny and Miles (1983)]. Such *service innovations* developed and diffused through the OECD countries in the last three decades have contributed to the continued growth of the personal final services sector.

It is being suggested that this dual economy comprising of a formal highly efficient, and internationally competitive sector for the production of material goods, and an informal, labor-intensive, low-wage sector which produces final services traded on a small scale may be a harbinger of things to come [Gershuny and Miles (1983)]. The possibilities of extending this model to a wide variety of final non-marketed services – entertainment, information, education and medical – look promising when viewed against recent developments in telecommunications, computing, information storage and retrieval technologies. A variety of innovative services can be provided comprising of

- new manufactured products in electronics, heavy electrical equipment and communications,
- a variety of intermediate services (computer programs, entertainment, educational, health care and counseling software, equipment maintenance, etc.), and
- new telecommunications infrastructure linking production sites, service delivery centers, households and community centers.

The application of innovations, for example, in fields such as community or home based long-term elderly care will involve major reorientations of the modes of provision of these services.

4.2. Rigidities in structural adaptation

Several recent analyses of the future prospects for growth and change in the industrialized world have drawn attention to

- certain classes of problems that increase the rigidities in adaptation to change, and
- certain trouble spots in existing institutions that suggest reduced capacity to adapt to change [this section draws heavily on OECD (1981)].

Four problem areas that augment rigidities are the following.

(1) *Demographic evolution.* The aging of the population leads to decreased occupational and geographic mobility, increasing dependency ratios (needs for costly services) as well as attitudes less congenial to structural adaptation.

(2) *Labor market.* The recent increase in 'structural' type of unemployment derives (besides the growth of the labor force) from long-term trends such as

(a) the mismatch between the pattern of low wage job supply (increasing flextime, or part time jobs) and the attitudes and skills of job seekers (better educated, lowered occupational mobility, etc.), and
(b) the creation of rigidities by government policies (varying between countries in the OECD) in the areas of social protection and labor costs. Such policies have increased overall costs, reduced the supply of jobs and saddled certain categories of persons with the burden of unemployment.[4]

(3) *Patterns of state intervention.* Certain rigidities are introduced into industrial societies by some forms of state intervention. Adjustments may be necessary in order to accommodate the changing environment described above. The two relevant aspects of state intervention pertain to

(a) many rigid regulatory policies (which have been necessitated by social concerns resulting from many forms of market failure). It is preferable to have policies that modify individual behavior while maintaining decentralization of the decision process would permit more continuous and flexible adjustment – hence the current demand for deregulation in many areas,
(b) public expenditure patterns: while the overall growth in the level of public expenditures is being resisted, potentially major conflicts are also rising in the allocation of expenditures among different categories (e.g., defense, welfare, debt servicing, infrastructure, etc.) and different social and economic groups.

(4) *International trade.* Increasing competition among developed countries and between them and some rapidly industrializing developing countries is strengthening moves to neoprotectionism in the form of import restraints and regulation of the international market for certain industries. Such efforts at neoprotectionism hamper structural adjustment in industrialized countries and the burden of adjustment is transferred to other countries.

While the above four problem areas pose increasing challenges to structural adaptation, the capacity to adjust in our institutional make-up shows also some 'trouble spots'. Examples are provided by

– *social oligopolization* – the organization of various interest groups (e.g.,

[4]Certain forms of social protection are increasing considerably the marginal cost of labor while reducing labor mobility and incentive to work. However, other rigidities are closely connected with social justice. A choice may be necessary between the immediate benefits of redistribution and indirect consequences even in terms of social justice, of economic inefficiency [OECD (1981, p. 170)].

doctors, farmers, businessmen, labor, etc.) in pluralistic democratic societies, in order to push for their respective viewpoints, leads to increased rigidities in international cooperation, inflation control, deregulation, etc.;

- *the faltering performance of the market and the welfare state* – the two major institutions for the distribution of goods and services; a major current issue is to use both these institutions in a manner that gets the best out of each, improve the functioning of each and possibly fashion an associative non-market, non-government sector for delivery of certain classes of services;
- *the problem of political institutions* – the key issues here are inefficiencies, 'overgrowth' of many government agencies, and the conflicts resulting from demands for decentralization and participation.

Finally in this necessarily cursory review of the sources of rigidities, it must be noted that these characteristics vary from country to country. Countries with certain industrial profiles (e.g., with shipbuilding, textiles, steel, etc.) are more vulnerable to international competition. In other cases, similar production structures make countries less complementary than before. Further, the labor markets in Japan and the U.S.A. appear to be more adaptable than in many European countries.

4.3. Implications for urban evolution

We present here some highly speculative observations on the impacts of this broad range of anticipated changes, and their pressures for structural adaptations, on urban activities and their organization in space.[5]

While the impacts of such broad changes may be far reaching (in the sense of rapid changes affecting large parts of the population) on individuals, the consequences on large organizations and social systems (e.g., urban form, the household economy, etc.) may be evolutionary. Different decision units evidence varying speeds of response. Productive capital, sensitive to a changing environment for opportunities, may respond in a short period. Human capital takes a larger response time. Housing capital turns over more slowly and infrastructure capital even more slowly. The speed of evolution of different forms of capital and activities in urban areas will very much depend upon the role of the public sector (discussed in section 5).

It may be useful to distinguish between the various urban impacts from two perspectives. The first type pertains to the impacts on activities that are urban in location and the second relates to impacts on national economic

[5]Noteworthy explorations of effects of the emerging information driven changes in the work place on future urban form appears in Simon (1980), Brotchie, Nijkamp et al. (1985), and Beaumont and Keys (1982).

activities that can occur in either urban or rural space [Lakshmanan and Chatterjee (1977)].

The essentially urban activities relate to three elements of the urban physical environment; shelter for residential and production activities; transport and communication facilities to link up production sites and residences, and a variety of facilities for the provision of public (or semi-public) services. Corresponding to these elements of the urban built environment is a complementary management or control system comprising of skilled individuals, organizations and institutions, that provides the knowledge and control base. Our earlier discussion, in section 2, of the demographic, transportation, industrial and residential evolution underway for the last two decades suggests increased spatial choice in the location of all the facilities in the urban built environment. The emerging information-driven changes in transportation, communications, production and service delivery is likely to lead to further enrichment of choice e.g., demand activated personalized transport, decentralized activities in the service sector, etc. As the traditional constraints of space and time are progressively loosened by improved decentralized services (e.g., electronic cottage, video conferencing) there may be alterations of timing and of activities. Thus a combination of structural trends (e.g., changes in the organization of production and service delivery, organization of the world economy, life styles and human skill evolution), provide strong forces working towards dispersed urban patterns. As a consequence, the recent trend for a highly differentiated, geographically dispersed structure of centers and subcenters will likely be accentuated.

However, there are also forces at work that promote regional concentration of economic activities: e.g., the need for face to face contact in rapidly evolving (highly information oriented) activities; the secular trend towards higher energy costs; the 'pull' of existing public investments; the attraction of the centers of new technology development; the pull of recent gentrification in central parts of the cities.

What is the likely outcome of the interplay of these opposing tendencies? One hint can be gained from the inertia of the existing built environment and the organizational forces guiding it which are likely to dampen any radical spatial reordering. Consequently, one can speculate that the emerging spatial pattern will be a diversified but integrated pattern of centers and subcenters with complex linkages among them. Perhaps a more reliable approach is to examine the adjustments that the anticipated changes call for from various national activities (whether in urban or in rural locations); for such adjustments collectively define the future urban form. The second class of urban adjustments are discussed next.

A central feature of the anticipated structural changes is the dynamic disequilibrating processes at work. While there is no doubt of certain elements of continuity in the economic process, change is often rapid in many areas. As new products are introduced new methods of production appear; new markets and new sources of supply emerge as the process of continuing disequilibria gets underway.

Two outcomes ensue. On the one hand, pressure builds up for increased mobility of factors of production to reflect the changing economic environment. Demand for capital and labor in new sectors and regions arises. Demand for new skills and new materials emerges. Thus there are clear benefits to the economy from increasing factor mobility. On the other hand, machinery, plants and labor in some industries and locations undergo devalorization. Residential capital owned by labor in sectors and locations suffering decline are particularly vulnerable to this devalorization. Consequently, there are serious costs to segments of the community arising from this demand for greater factor mobility. Thus there is likely to be organized pressure against shifting capital rapidly out of economic activities and regions with declining prospects.

The dialectic between the pressures for and resistance to increased factor mobility determines the pattern of economic adjustments in terms of the speed of growth of new activities, the ordered decline of other sectors, and the changing fortunes of residential capital in specific locations. It is on such an outcome that the future geography of economic activities will depend. In such an outcome, the role of the public sector in tilting the balance one way or the other is clearly crucial. It is to a clarification of the role of the public sector we turn next.

5. Implications for urban policy

A major argument running through this paper is that the rate and direction of future technical change and adjustments to it in terms of the organization and location of urban activities cannot be separated from the broader economic and political setting within which they take place. For example, the cost and feasibility of technical changes in energy extraction, in communications infrastructure, and in materials usage will be significantly affected by the environmental, occupational safety and social legislation in effect. The potential ability to provide the broad enrichment of choice for producers and consumers in the emerging information-rich era will depend upon the institutional and organizational responses; broadly speaking, the pace and direction of technical and organizational innovations will be powerfully influenced by the system of incentives in place. The manner in which such a system of rewards and penalties operates to promote technical change and adjustments to it will depend to a considerable degree on the nature and forms of public sector intervention.

Various scholars have referred to the important role of the public sector in the major transformations [Polyani (1944), De Brunhoff (1978)]. The areas and forms of public intervention have varied, however, with the conditions and requirements of the period. The public sector played a crucial role in the pursuit of commercial capitalism through its 'economic policy' of mercantilist regulations affecting markets. In early industrialization, the role of the public sector was to institute the dominant laissez-faire system and to facilitate the mobilization of capital. As the factory system developed, the economies of scale and external economies led to concentration of capital and economic activities in space with a broad range of labor dislocations and environmental externalities. In further development, the burden of adjustment to mobility of capital fell on certain groups and regions depending on their factor endowments. There was, on the part of the state, no anticipation of the adjustments that households or work force had to make to these wrenching changes. As the breakage and wastage multiplied and social costs mounted, the public sector assumed an ameliorative role guided by notions of antimonopoly, social justice and welfare, instituting a variety of legislations pertaining to labor organization, minimum wage, welfare payments, safety and environmental quality with investments in public infrastructure.

Once more the conditions and requirements for public policy appear to be changing. While there is much room for intellectual pluralism about the nature of these changes, some relevant features of the emerging socio-economic context appear to be smaller, 'foot-loose' high technology units with a bimodal skill distribution (demand for highly skilled and low skilled labor but with limited demand for middle skill groups) and a big appetite for venture capital; a robust 'informal' sector in a variety of service activities with flexible time schedules; considerable changes in the organization of the household economy; an overall demand for greater mobility of capital and labor and for adaptive reuse of parts of the urban built environment; and all of this accompanied by a strong pressure for deregulation on a wide front. While the pressure to intervene in many areas of social protection is often unavoidable, the desired form should permit dynamism, increase innovative capacity and continuous monitoring.

What do these changes imply for a new 'posture' for the public sector? On the one hand, the public sector is likely to promote efficient adjustments to change through encouraging factor mobility, appropriate institutional innovations and a greater articulation of the local, national and international aspects of policy in a global economy. On the other hand, given the social consciousness of the significant costs of past dislocations, future public intervention can ill afford to be ex post and ameliorative. It should be anticipatory and 'developmental' with regard to the costs of adjustments of small producers and households.

What does such a 'posture' mean in terms of specific forms of public intervention? In our view, such forms involve, among other things, reduction of the costs of adjustment and hastening of the adjustment process through institutional innovations. For example, factor mobility can be promoted through measures of risk spreading. The demand for venture capital can be met by risk reduction through insurance and tax write offs. Indeed this is one area where practical endeavors have outpaced analytical research. Many states such as Massachusetts, Connecticut and New Jersey have fostered availability of venture capital long before the flowering of analytically focused research in the area [Bearse and Konopka (1979)].

In the area of human capital, the demand for increased technical education and a continuing education process would require fresh organizational and locational responses from educational institutions. In the urban built environment adaptive reuse can be fostered. Such reuse is exemplified not only by conversion of piano factories to residence workplaces as in Boston but also by the broader gentrification process in central parts of metropolitan areas. Adaptive reuse is actively promoted by various public sector incentives in the U.S. More challenging would be the imaginative reuse of residential capital stock left behind or devalorized by redundant relocating labor or by footloose groups such as the elderly.

Another likely form of public sector intervention is institutional innovation. To smooth the operation of the 'dual economies' developing in some segments of final marketed and non-marketed service sectors and to promote institutional competition among private and public service providers, a number of social experimentation efforts may emerge. Such experiments (e.g., the housing supply experiment, income maintenance experiment, etc.) help identify efficient innovations for the delivery of final market and non-market services.

The discussion on forms of likely future public intervention has a twofold relevance to students of metropolitan dynamics. First, future outcomes in the course of metropolitan evolution are heavily influenced by such public interventions. Second, because of their far reaching effects, specific forms of intervention will undergo ex ante assessment of their consequences on desired social objectives. Consequently, many of the policies considered here are components of metropolitan modeling.

6. The analytical agenda for metropolitan modeling

Much of urban modeling, at its inception a quarter century ago, was very much a pragmatic endeavor, located in urban planning and transportation studies [e.g., Hansen (1959), Lowry (1964), Lakshmanan and Hansen (1965)]. The emphasis was on operationality and management of data, and not on logical deductive systems. As the field matured, there was greater formalism as in the generalization of the spatial interaction model through notions of entropy maximization by Wilson (1974) and others. The NBER model of the urban housing market and transport system developed by Kain and Ingram

(1976) illustrates the microeconomic tradition. Recent theoretical developments in the 'new urban economics' and discrete choice theory are being incorporated into urban modeling. Further the tradition of partial equilibrium modeling dealing with well-defined subsystems of the metropolis is being modified by recent work on integrated modeling of e.g., I–O and demographic models [Gordon and Ledent (1980), Batty and Madden (1981)], urban and transport models [Echenique (1977)] and a full range of urban activity models [Johansson and Snickars (1985), Brotchie, Dickey and Sharpe (1980)].

However, in all of this flowering of sophisticated modeling, the focus on relevant substantive questions – what issues are to be modeled – is rather limited. As Batty (1983) notes, 'one of the great disappointments of modeling practice has been the inability of the theorists to suggest model structures which capture the qualitative flavor of urban systems and problems'. While the progress of modeling techniques has been impressive, the system being modeled has remained reasonably stable, in spite of the remarkable shifts in the perceptions of what urban planners and policy makers view as relevant.

The thrust of this paper is to highlight such questions relevant to future metropolitan evolution that need to be embodied in mathematical models of the urban system. Such questions will form the analytical agenda of models intended to capture metropolitan dynamics. For purposes of exposition, we group such issues into four classes. As we present these issues below, we also argue that their incorporation has significant implications on future 'model styles' as well. We will draw attention to one of these styles of metropolitan modeling.

The classes of issues that are crucial to the substantive agenda of dynamic metropolitan models in addition to those usually addressed are

- dynamic adjustment processes,
- distributional issues,
- national links of the urban system,
- modeler–decision maker interface.

6.1. Dynamic adjustment processes

As argued so far, traditional analyses of comparative statics are inadequate for the task at hand. Metropolitan processes of adjustment to a variety of changes need to be modeled. An understanding of the nature and incidence of the costs incurred in this adjustment is both intellectually interesting and vital for the implementation of the transition policies. It is useful to distinguish between three types of adjustment processes:

- short term (up to one year),
- medium term (2–5 years) in which new plants can be built, households can

relocate, new technological processes be designed and built, new sources of raw materials can be developed, new legislation passed, and new supportive public investments initiated,

- long term (10–20 years) when plant, equipment and some types of capital stock become obsolete, significant changes in an urban infrastructure and the built environment are required.

The response times to change will clearly differ markedly with industrial sector, type of household and public agency. The various economic, behavioral and technological relations incorporated in urban and regional models will vary for each dynamic adjustment process. Fixed model parameters in the short term become variables in longer adjustment periods. Since economic, social, technological and urban development phenomena in the real world embody such dynamic adjustment processes, satisfactory models must incorporate them for the time periods over which they are believed to be appropriate; otherwise conditional or policy analysis predictions made from the model will be grossly inaccurate.

Econometric and I–O models obtained by fitting historical data implicitly assume that the estimated parameters are constant. Mathematical programming, based on engineering estimates of processes, can incorporate technological processes at future points in time, but they do not easily make the transition from the present to the future. Some economic models distinguish between short- and long-term price elasticities, while others let I–O coefficients vary with relative prices [Hudson and Jorgenson (1974)]. While these developments are relevant, they are only a partial treatment of structural change processes occasioned by technical changes which are likely to lead to changing technological and industrial mixes, and the location of population and production.

Some recent modeling efforts have begun to address this problem. The ideas of Berndt et al. (1979) in developing dynamic adjustment models of industrial demand have been applied at the regional level [Lakshmanan (1982), Anderson and Lakshmanan (1983)] and could be easily incorporated into regional/national models. Another example is the specification of adjustment processes via vintage modeling techniques [Persson and Johansson (1982)]. Thus there is a potential for better specification of economic adjustment processes than is currently the case at the metropolitan levels.

At the intrametropolitan level, knowledge of the dynamics (interaction of variables and speed of adjustment) in models of employment, housing, and land use is primitive. Generally, the interactions of variables are myopic (with current decisions entirely dependent on the outcomes of current parameter values) and adjustments are close to instantaneous. At the local level, dynamic processes of stock adjustment have not yet been analytically explored. Housing stock turns over slowly; productive and human capital much faster, leading to stock mismatches and disequilibria, with potential impacts on the quality of the urban environment. Analytical improvements in this area are a research frontier in urban and metropolitan modeling.

Most of the models dealing with the economy or the urban environment are derived from a neoclassical framework that does not pay much attention to technological innovation;[6] yet dynamic processes such as innovation, innovation diffusion, and capital investment are obviously part of the adjustments to technological developments [Nelson and Winter (1983)]. Behavior-oriented models of households and firms that depict information generation flows, communications, and processing are needed, and could lay the foundation for the understanding necessary to identify dynamic adjustment processes in the various components of a metropolitan system.

6.2. Distributional issues

New production, service and information technologies will lead to an uneven distribution of benefits and costs – depending upon the ownership of assets and skills that are being valued or devalorized, or location in areas visited by rapid growth or decline. It is important to take such distributional outcomes into account and to incorporate them in politically responsive planning processes. Such elements will become even more important in urban analysis if economic growth rates of developed countries slow down considerably.

It is here that the recent microsimulation modeling tradition in the U.S.A. is relevant [Orcutt et al. (1976)]. These models developed in the U.S. Poverty Institute and *Mathematica*, utilize recently available data on household behavior to provide information on labor supply, savings, consumption migration, and other responses to external change and public policy by income class. Another class of distributional effects is spatial and is incident on local government jurisdictions due to the changing distribution of households and enterprises resulting from large-scale technical changes. The consequent interjurisdictional shifts in public revenues and expenditures are of interest to local decision makers and also need to be identified.

6.3. National linkages

Given the emerging strong two-way flows of information and factors of production between a metropolitan region and the nation, these links become objects of analysis. The traditional 'top down' formulation of these links need to be supplemented by 'bottom up' treatment of certain variables as some types of investment.

[6]A model system deriving from a different perspective – that of self organizing systems – by Allen et al. (1981) addresses issues of urban dynamics such as disequilibria in terms of demand and supply, discontinuity, and threshold.

The partial equilibrium orientation has persisted in urban modeling in spite of the oft repeated qualification that general equilibrium effects matter or 'everything is related to everything else'. Thus one frequently encounters some form of fixed coefficient, reduced form models of demographic and economic changes in urban areas. It is here that the recent efforts at computable general equilibrium (CGE) modeling may be relevant. These multisector models center on factor movements between sectors and across space as well as on the role of wage and price signals to guide this transformation. Metropolitan growth not only lends itself to CGE modeling but CGE may offer a modeling style for studying economy/urbanization issues [Kelley and Williamson (1984)].

6.4. Modeler–decision maker interface

In urban modeling, the interface between the producers and users of models has been a bothersome barrier to its success. It may be useful to note here that differences in knowledge, functional roles, and time scales of operation between modelers and decision makers account for this barrier. Further, there is an essential tension between scientific policy analysis (deriving largely from a rationalistic goal oriented framework) and decision making in democratic pluralistic societies.

Policy issues in metropolitan contexts tend to be complex, ill-structured, and amorphous. Model systems are rather monolithic in the sense that the overall objective function determines optimality, while the real world has pluralistic decision structures, is interactive, and full of conflicts. These models thus tend to be useful in centralized decision contexts (as in typical operations research contexts). More often, in the public policy arena, economic modeling is inadequate in that it does not yet provide a method for strategic analysis that allows, in a context of advisory procedures and bargaining, an orderly approach to balancing model results and judgment.

It is in this context that one can understand the strong demand for interactive multilevel modeling [Despontin, Nijkamp and Spronk (1983)]. While a few such modeling procedures have been sketched in carefully delimited contexts, their potential in the complex, ill-defined arena of urban decision making is still limited for enhancing the mutual understanding of the issues and problems among modelers and the multilevel participants in decision making. Developments in decision analysis, to sharpen this process of mutual understanding and learning, are sorely needed.

References

Allen, P.M., M. Sanglier, F. Boon, J.L. Deneuborg and A. DePalma, 1981, Models of urban settlement and structure as dynamic self-organizing systems (U.S. Department of Transportation, Washington, DC).

Alonso, William, 1978, The current halt in the metropolitan phenomenon, in: Charles L. Leven, ed., The mature metropolis (Heath, Lexington, MA).

Anderson, W. and T.R. Lakshmanan, 1983, A model of regional dynamic adjustment in U.S. manufacturing, Working paper (Department of Geography, Boston University, Boston, MA).

Batty, Michael, 1979, Paradoxes of science in public policy: The baffling case of landuse models, Sistemi Urbani 1, 89–122.

Batty, Michael, 1983, Technical issues in urban model development, Papers in planning research (Department of Town Planning, University of Wales, Cardiff).

Batty, Michael and M. Madden, 1981, Demographic economic forecasting within an activity–commodity framework, Environment and Planning A 13, 1067–1083.

Bearse, Peter J. and D.A. Konopka, 1979, A comparative analysis of state progress to promote new technology based enterprise, The New England Journal of Business and Economics 5, no. 2, 49–74.

Beaumont, J.R. and P.L. Keys, 1982, Future cities: Spatial analysis of energy issues (Wiley, New York).

Berndt, E.R., M.A. Fuss and L. Waverman, 1979, A dynamic model of costs and adjustments and interrelated factor demands, with an empirical application to energy demand in U.S. manufacturing, Discussion paper no. 79-30 (Department of Economics, University of British Columbia, Vancouver).

Briggs, Asa, 1979, The age of improvement 1783–1867 (Longman, London) Chapters 1 and 10.

Brotchie, John, Peter Nijkamp et al., eds., 1985, Technical change and urban form (Croom Holm, London) forthcoming.

Brotchie, J.D., J.W. Dickey and R. Sharpe, 1980, TOPAZ – general planning technique and its applications at the regional urban and facility planning levels, Lecture notes in economics and mathematical systems 180 (Springer, Berlin).

Browning, H.C. and J. Singlemann, 1978, The transformation of the U.S. labor force: The interaction of industry and occupation, Politics and Society 8, 481–509.

Channon, Derek F., 1978, The service industries (Macmillan, London).

Clark, Colin, 1940, The conditions of economic progress (London).

Cross, Nigel, David Elliott and Robin Roy, 1974, Man-made futures: Readings in society, technology and design (Hutchinson, London).

De Brunhoff, Suzanne, 1978, The state, capital and economic policy (Pluto Press, London).

Despontin, W., P. Nijkamp and B. Spronk, 1983, Macroeconomic and multiobjective modeling (Springer-Verlag, Berlin).

Echenique, M., 1977, An integrated land use and transport model, Transactions of the Martin Center 2, 195–230.

Fisher, A.G.B., 1935, Economic implications of material progress, International Labor Review.

Flinn, M.W., 1966, The origins of the industrial revolution (Longman, London).

Freeman, Christopher, 1982, The economics of industrial innovation (Frances Pinter, London).

Friedrichs, Gunter and Adam Acheff, eds., 1983, Microelectronics and society: A report to the club of Rome (Pergamon Press, London).

Fuchs, V.R., 1968, The service economy (National Bureau of Economic Research, New York).

Gershuny, Jonathan, 1978, After industrial society (Macmillan, London).

Gershuny, J.I. and I.D. Miles, 1983, The new service economy (Frances Pinter, London).

Gordon, P. and J. Ledent, 1980, Modeling the dynamics of metropolitan areas: A demoeconomic approach, Environment and Planning A 12, 125–133.

Hansen, Walter, 1959, Accessibility and landuse, Journal of the American Institute of Planners (special issue on the nature and use of models in city planning, Alan M. Voorhees, ed.).

Hartwell, R.M., 1967, The causes of the industrial revolution in England (Methuen, London).

Hartwell, R.M., 1971, The industrial revolution and economic growth (Methuen, London).

Hudson, E. and D. Jorgenson, 1974, U.S. energy policy and economic growth, 1975–2000, Bell Journal of Economics and Management Science 5, 461–514.

Johansson, B. and F. Snickars, 1985, The Swedish case study, in: T.R. Lakshmanan and B. Johansson, eds., Large-scale energy projects: Assessment of regional consequences, (North-Holland, Amsterdam).

Kain, J. and G. Ingram, 1976, The NBER model (Harvard University Press, Cambridge, MA).

Kelley, Allen C. and Jeffrey G. Williamson, 1984, What drives third world city growth? (Princeton University Press, Princeton, NJ).

Klein, B.H., 1977. Dynamic economics (Harvard University Press, Cambridge, MA).
Kondratiev, N.D., 1935, The long waves in economic life, Review of Economic Statistics, Nov.
Lakshmanan, T.R., 1982, Integrated multiregional economic modeling for the U.S.A., in: B. Issaev et al., eds., Multiregional economic modelling: Practice and prospect (North-Holland, Amsterdam) 171–188.
Lakshmanan, T.R. and Lata Chatterjee, 1977, Urbanization and environmental quality, Resource paper 77-1 (Association of American Geographers, Washington, DC).
Lakshmanan, T.R. and Walter G. Hansen, 1965, A market potential model, Journal of the American Institute of Planners (special issue edited by Britton Harris).
Landes, D., 1969, The old bank and the new: The financial revolutions of the 19th century, in: F. Crouzet, W.H. Chaloner and W.M. Stern, eds., Essays in European economic history 1789–1914 (Longman, London) 112–127.
Leven, Charles L., ed., 1978, The mature metropolis (Heath, Lexington, MA) 3–20.
Levinson, Irving and J.W. Wheeler, eds., 1981, Western economics in transition, Hudson Institute studies (Westview Press, Boulder, CO).
Lowry, Ira S., 1964, A model of metropolis, RM-4035-RC (The Rand Corporation, Santa Monica, CA).
Mensch, G., 1979, Stalemate in technology (Ballinger, Cambridge, MA).
Nelson, R.R. and S.G. Winter, 1977, In search of useful theory of innovation, Research Policy 6, 36–76.
Nelson, R.R. and S.G. Winter, 1983, An evolutionary theory of economic change (Harvard University Press, Cambridge, MA).
OECD, 1978, A medium term strategy for employment and manpower policies (Paris).
OECD, 1981, Interfutures (Paris).
Olson, M., 1982, The rise and decline of nations (Yale University Press, New Haven, CT).
Orcutt, G.H. et al., 1976, Policy exploration through microanalytic simulation (The Urban Institute, Washington, DC).
Perloff, Harvey S., 1978, The central city in postindustrial age, in: Charles L. Levin, ed., The mature metropolis (Heath, Lexington, MA) 109–129.
Persson, H. and B. Johansson, 1982, A dynamic multisectoral model with endogenous formation of capacities and equilibrium prices: An application to a Swedish economy, Professional papers 82–89 (IIASA, Laxenburg).
Polyani, Karl, 1944, The great transformation (Rinehart, New York).
Rothwell, Roy and W. Zegveld, 1981, Industrial innovation and public policy (Frances Pinter, London).
Rothwell, Roy and W. Zegveld, 1982, Innovation and the small and medium firm (Frances Pinter, London).
Schaffer, Bernard and Geoff Lamb, 1981, Can equity be organized? (Gower, London).
Schmookler, J., 1966, Innovations and economic growth (Harvard University Press, Cambridge, MA).
Schumpeter, J.A., 1939, Business cycles (McGraw Hill, New York).
Simon, J., 1980, The consequences of computers for centralization and decentralization, in: M.A. Dertonzous and J. Moses, eds., The computer age (The MIT Press, Cambridge, MA).
Thompson, Wilbur, 1965, A preface to urban economics (The Johns Hopkins University Press, Baltimore, MD).
Utterback, J.M. and W.J. Abernathy, 1978, Patterns of industrial innovation, Technology Review 80, no. 7.
Wigan, M.R. and J.M.M. Morris, 1980, The transport implications of activity and time budget constraints, Transport Research 15A, no. 1, 56–66.
Wilson, Alan G., 1974, Urban and regional models in geography and planning (Pion, London).
Wolman, W., 1983, Central government policies toward declining urban economies, Environment and Planning A 15, 300–315.
Zwan, A. van der, 1979, On the assessment of Kondratiev cycle and related issues, Center for Research in Business Economics, Mimeo. (Rotterdam).

Regional Science and Urban Economics 16 (1986) 31–42. North-Holland

THE EVALUATION OF COMPLEX URBAN POLICIES

Simulating the Willingness to Pay for the Benefits of Subsidy Programs*

John M. QUIGLEY

University of California, Berkeley, CA 94720, USA

Received October 1984, final version received March 1985

This paper considers the evaluation of urban renewal and other urban investment policies and the application of the aggregate willingness to pay criterion to investment decisions. Two rigorous approaches to the measurement of program benefits are examined. The two methods, 'hedonic pricing' and 'quantal choice', are compared by relying on a series of simulations.

1. Introduction

To an economist, the justification for publicly provided urban shelter or urban renewal projects is much stronger if they are socially 'profitable' than if they are merely politically acceptable forms of redistribution towards deserving groups. Indeed, if such programs pass the cost–benefit criterion of social profitability, then efficiency can be improved with no sacrifice of distributional goals. Urban shelter or urban renewal projects can be justified as socially profitable, in turn, if either suppliers or demanders face inefficient price signals – for example, if some prisoners' dilemma prevents atomistic suppliers from maximizing collective profits or if some non-marketed external benefit incidental to housing consumption prevents demanders from maximizing utility.

To justify some urban shelter program on the basis of consumption externalities, it would be necessary to demonstrate that an urban housing investment program provided increased public health, safety or labor market benefits which were not reflected in consumers' willingness to pay for increased amenities. Although there is no lack of assertions about the

*A previous version of this paper was presented at the IIASA Conference on Metropolitan Processes and Policies in Rotterdam, June 1984. The paper benefited from the comments of participants at Berkeley's seminar on urban economics and local finance and from discussions with Pravin Varaiya. It could not have been produced without the expert assistance of Carl Mason, who programmed many of the calculations and who uncovered a serious error in a previous draft.

0166–0462/86/$3.50

importance of these externalities, there is little or no serious evidence to support their existence.[1]

In contrast, to justify an urban shelter or urban renewal project due to failures in individual profit signals, it would be necessary to demonstrate that the aggregate willingness to pay for improved conditions exceeds the supply cost of provision. This criterion is quite demanding. Nevertheless, casual empiricism suggests that many proposed projects might satisfy it. Neighborhood externalities and the propinquity of parcels, buildings, and land uses all seem to indicate that the level of private and public investment in urban amenities could deviate substantially from the level that would maximize individual well-being and collective profits.

This paper considers the evaluation of urban renewal and other urban investment policies and the application of this aggregate willingness to pay criterion to investment decisions. In the following section, we discuss two rigorous approaches to the measurement of program benefits, each deduced from general theoretical notions advanced a decade ago. In section 3, we compare the two methods by relying upon a series of simulations. The preliminary results are presented in section 4.

2. Willingness to pay for urban shelter programs

Consider the following general problem. A household of income y is observed to consume a vector of housing and urban amenities at some market price (e.g., monthly rent). Without loss of generality, assume the vector of amenities is h at market price $p(h)$ leaving y minus $p(h)$ for the consumption of other goods. As a result of some public investment policy, the household is offered the opportunity to consume H at some price $p(H)$. What is the household's willingness to pay for the public program?

Let U (.,.) be the utility function for the household. The amount of money Δ which could be given to the household in lieu of the proposed public investment program is the solution to

$$U(h, y-p[h]+\Delta)=U(H, y-p[H]). \tag{1}$$

The amount of money, δ, which could be taxed from the household benefiting from the public program to leave it as well off as it was initially is the solution to

$$U(H, y-p[H]-\delta)=U(h, y-p[h]). \tag{2}$$

[1]See Burns and Grebler (1976), for example, for a critique of such evidence from developing countries where, it has been asserted, the magnitude of such external consumption benefits is large.

These measures, the so-called Hicksian equivalent variation and compensating variation, represent the cash value of the public program in terms of its effects upon the household. In the absence of general equilibrium effects, the aggregate willingness to pay for the investment program is merely the sum of Δ or δ over the relevant population.

The estimation of these magnitudes requires some theoretically defensible procedure for inferring the 'shape' of the ordinal utility function or for deducing the compensated demand curve for the vector of amenities. Extensions of Rosen's work on hedonic prices and McFadden's work on discrete choice models, both published in 1974, provide alternative methods for using market information to estimate utility contours rigorously and for measuring the benefits of public investment programs rigorously according to eq. (1) or (2). In the remainder of this section, we indicate how these models can be applied to estimate program benefits.

2.1. A continuous model: Hedonic pricing

Two of the distinguishing features of the housing market are that a large fraction of the housing services consumed in a given period is produced from the standing stock and that the stock is itself expensive to modify. Thus, to a first approximation, housing prices are demand determined, as existing dwellings are 'auctioned' for occupancy by the highest bidder in any period. In addition the resulting housing prices are generally non-linear functions of quantity, due to high transformation costs (or 'repackaging' costs, in Rosen's terminology).

Utility maximization implies that each household chooses h to

$$\max U(h, y-p[h]) \tag{3}$$

with a given exogenous price function. The price function itself is determined endogenously from the competitive behavior of households solving the above maximization problem. This hedonic price function is given by the solution to

$$\frac{\partial U(h, y-p[h])/\partial h}{\partial U(h, y-p[h])/\partial y} = \frac{\mathrm{d}p}{\mathrm{d}h}. \tag{4}$$

The left-hand side of (4) is the marginal rate of substitution of housing for income, the income compensated demand for housing, or the household's marginal bid for an additional unit of housing. In equilibrium, the marginal bid just equals the marginal price of housing dictated by the market, the right-hand side of eq. (4).

Given an exact functional form for the utility function, and given some mapping of housing to income, $y=f(h)$, the market-wide hedonic price relationship can be computed.

For example, if the utility function is Cobb–Douglas with parameters α and β,

$$U=Ah^{\alpha}(y-p[h])^{\beta}=Ah^{\alpha}(f[h]-p[h])^{\beta}, \tag{5}$$

then integration of (4), with initial condition $p(1)=0$, yields

$$p(h)=(\alpha/\beta)h^{-\alpha/\beta}\int_{1}^{h} f(u)/u^{1+\alpha/\beta}\,\mathrm{d}u. \tag{6}$$

If the mapping is known exactly, say the linear and continuous function

$$y=f(h)=h-1, \tag{7}$$

then the hedonic function can be derived explicitly,

$$p(h)=\{\alpha/(\alpha+\beta)\}h+\{\beta/(\alpha+\beta)\}h^{-\alpha/\beta}-1. \tag{8}$$

For different assumptions about the form of the utility function and for different mappings relating the distribution of housing to income, eq. (4) can be solved for the market-wide hedonic price relation. Of course, for plausible utility functions it may not be possible to solve (4) in closed form. The hedonic relationship between $P(h)$ and h may, however, be approximated quite easily using numerical methods. In a demand determined world, the exact locus of the hedonic function can be inferred from knowledge of the utility function and the housing–income frequency distribution.

Of course the essence of the welfare economics problem is that the parameters of the utility function are not known. They must be inferred from limited information about market behavior. In contrast, the hedonic price relationship can be 'observed' directly in a market, at least by statistical means. In particular, a body of observations on dwellings and their prices permits the computation of some regression approximation to the 'exact' hedonic function. This, in turn, permits statistical estimation of the parameters of an assumed functional form for the utility function. For example, if the form of the utility function is GCES,

$$U=[\alpha h^{\beta}+(y-p)^{\varepsilon}]^{\phi}, \tag{9}$$

then substitution into (4) yields

$$\frac{\alpha\beta}{\varepsilon}\frac{h^{\beta-1}}{(y-p)^{\varepsilon-1}}=\frac{\mathrm{d}p}{\mathrm{d}h}, \quad \text{or} \tag{10}$$

$$\log(\alpha\beta/\varepsilon)+(\beta-1)\log h+(1-\varepsilon)\log(y-p)=\log(\mathrm{d}p/\mathrm{d}h). \tag{11}$$

The parameters of the utility function, α, β, ε can be estimated consistently from the three coefficients of the regression estimate of eq. (11). The dependent variable is the logarithm of the derivative of the hedonic price function and the independent variables include the logarithms of the consumption of housing and other goods. If the utility function is CES, $\theta=\alpha=\beta=\varepsilon=1/\phi$, then (11) can be simplified to

$$\log\theta+(\theta-1)\log(h/[y-p])=\log(\mathrm{d}p/\mathrm{d}h). \tag{11'}$$

2.2. A discrete model: Quantal choice

Another distinguishing feature of the housing market is the discrete nature of consumer choice. Although the housing bundle is composed of a large number of diverse components, housing choice consists of the selection of one unit out of a potentially large number of discrete alternatives. In this market, a household chooses a specific and discrete dwelling to solve the maximization problem in (3). In particular, as McFadden (1974) has shown, if the individual utility function includes an additive stochastic component, and if the stochastic component is independently and identically Weibully distributed across households, then the probability, Π, that a household will choose a particular dwelling, h^*, is

$$\Pi(h=h^*)=\exp\{U(h^*,y-p[h^*])\}\Big/\sum_h\exp\{U(h,y-p[h])\}. \tag{12}$$

If the preference function is linear in parameters, then these parameters may be uniquely estimated, up to a factor of proportionality, by maximizing a log likelihood function of the form

$$\log L\propto\frac{1}{k}\sum_k\frac{\log \mathrm{e}^{U(h^*,y-p[h^*])}}{\sum_h\log \mathrm{e}^{U(h,y-p[h])}} \tag{13}$$

for a sample of k observations on choices h^* and available alternatives h. Clearly the set of alternative dwellings in a metropolitan area is so large as to make an iterative solution of (13) computationally infeasible. However, as McFadden (1978) has shown, it is possible to estimate the choice model in a consistent manner by selecting a sample d of rejected alternatives for each household according to the sampling rule, Θ,

$$\text{if}\quad \Theta(d|h^*)>0,\quad \text{then}\quad \Theta(d|h^*)=\Theta(d|h). \tag{14}$$

This sampling rule posseses the so-called 'uniform condition property'. For each individual, the sample includes the chosen alternative, and each alternative in the set d is equally likely to be the chosen alternative. Under these conditions, the summation in the denominator $\sum_h$ can be replaced by $\sum_d$ and the parameters of the model can be estimated by maximum likelihood using a sample of metropolitan housing alternatives.

2.3. Housing market applications

During the past few years, there have been an increasing number of applications of these techniques to the housing market. Empirical analyses exploiting the non-linearity of housing prices to estimate the benefits of urban amenities have been reported by Harrison and Rubinfeld (1978), Kaufman and Quigley (1984), Quigley (1982), Witte et al., (1979), among others.

Empirical applications of the quantal choice model to the housing market include papers by Anas (1980, 1985), Ellickson (1981), Kain and Apgar (1977), Lerman (1977), Quigley (1983, 1985) and Williams (1979).

In applying these very different techniques to estimating demands for amenity, researchers have utilized the same kinds of data – a sample of households, their incomes and demographic characteristics on the one hand, and the characteristics of the dwellings these households occupy, including exogenous housing prices, on the other hand.

3. A stylized comparison of willingness to pay

Although they rely upon substantially the same data to answer similar questions about the slopes of consumers' utility functions, the hedonic and discrete choice approaches employ very different assumptions and statistical techniques. One objective of the analysis described in this section is to compare the implications of the two models using the same underlying data. In particular, a major objective of the comparison was to characterize the circumstances under which one or the other analysis is likely to provide more accurate estimates of the welfare benefits of programs. Currently, a more elaborate Monte Carlo comparison is under way to investigate how sensitive estimates of welfare effects are to changes in parameter values or to stochastic factors.

The second objective of the analysis is to consider explicitly the endogeneity of prices in the housing market and the effects of this endogeneity upon estimated welfare effects. As noted above, previous applications of these hedonic or discrete choice techniques to the housing market have assumed that housing prices are given exogenously. The comparative analysis in this section is based upon the market equilibrium prices determined by the choices of actors in the housing market.

3.1. The structure of the simulations

We conduct the simulations by choosing the form of the utility function for households in the market and the parameters of that function. Thus the compensated demand curves for housing and the shapes of the utility contours are known.

We next choose a continuous mapping from housing to income, $y=f(h)$, in the market. This is equivalent to selecting the joint frequency of income and available housing units. We assume throughout that housing is a normal good, that is, that the mapping is a monotonically increasing function.

The form of the utility function and the relative frequency distribution are sufficient to define the market clearing price relationship in the market, at least if housing is auctioned to the highest bidder. We assume that the market consists of 100 households of varying incomes and more than 100 dwellings. The structure of these prices is computed by numerical integration of (4) using the Runge–Kutta method. The price structure will have the following properties: all dwellings below those 100 which provide the highest levels of service (h) will be vacant. That dwelling numbering 100 from the highest in terms of h will be occupied at a price of zero, and the remaining 99 dwellings will be occupied at positive prices. The equilibrium price structure clears the market and assigns each household to its preferred dwelling. The endogenous price relationship represents the equilibrium pattern of housing prices in the market.

At this point a 'data set' has been created. The data set consists of 100 observations on households: their incomes y, their housing consumption h, their expenditures on housing, $p(h)$, and on other goods, $y-p(h)$. This data set is then analyzed using the two techniques described above: the so-called hedonic and quantal choice approaches. For each technique we estimate the parameters of the utility function, the slope of the contours, and we compare the estimates with the characteristics of the known function.

3.2. Estimating the hedonic model

We use the 100 observations on h and $p(h)$ to estimate the hedonic price function in the market by a power series approximation, i.e., we estimate the regression

$$p(h)=\omega_0+\omega_1 h+\omega_2 h^2+\omega_3 h^3+\omega_4 h^4+\omega_5 h^5=g(h). \tag{15}$$

We then differentiate this function, take logarithms, and estimate the regression

$$\log(\mathrm{d}g/\mathrm{d}h)=\eta_0+\eta_1 \log h+\eta_2 \log(y-p), \tag{16}$$

using the sample of 100 observations. The parameters of this regression are transformed to provide estimates of the GCES approximation to the unknown utility function,

$$U=\alpha h^{\beta}+(y-p)^{\varepsilon}, \quad \text{where} \tag{17}$$

$$\alpha=[(1-\eta_2)/(1+\eta_1)]e^{\eta_0}, \qquad \beta=1+\eta_1, \qquad \varepsilon=1-\eta_2.$$

3.3. Estimating the quantal choice model

For each of the 100 observations on the housing chosen by a household of income y at price p, we select a sample d, of four dwellings not chosen by the household according to the sampling rule

$$\Theta(d\,|h)=4/99, \tag{18}$$

that is for each household in the sample, we randomly select four dwellings which have been rejected according to a rule with the uniform conditioning property. We estimate the parameters of a linear approximation to the unknown utility function,

$$U=\gamma_1 h+\gamma_2(y-p)+\gamma_3 h(y-p), \tag{19}$$

from the observations on the chosen alternative and a sample of four rejected alternatives for each of 100 households. The estimation is undertaken by maximizing the likelihood function in (13) according to the procedure suggested by McFadden (1978).

4. Some preliminary results

This section presents a comparison of these methods of estimating the preferences of housing consumers using the same body of information. This information is, in turn, generated by a known structure of household preferences and some specified relative distribution of income and housing. The following example may provide a concrete illustration of the comparison.

Assume the structure of preferences is GCES with $\alpha=\beta=\varepsilon=0.250$ and there are 100 households with incomes, y, ranging from 1 to 11 in units of 0.1. Assume further that the market consists of 100 units whose quality level, h, is normally distributed with mean 5 and standard deviation 2. Housing is a normal good; the rectangular income distribution and the normal housing distribution yield a monotonic relationship,

$$y=(1/\sqrt{2}\pi)\exp[(h-5)^2/2], \tag{20}$$

between h and y in the market. Together these assumptions yield an equilibrium structure of prices (by integrating eq. 4) throughout the market.

Consider the hedonic approach. The equilibrium price structure is approximated by the continuous function in eq. (15), estimated by ordinary least squares. The derivatives of this function at the prices computed for each dwelling are then used to estimate the parameters of eq. (16) by ordinary least squares.

This procedure yields estimates of $\alpha=0.227$, $\beta=0.264$, and $\varepsilon=0.244$ for the three parameters, estimates which differ from the true values by 2 to 9 percent. Although the estimates differ from the true parameters, the values of the utility functions are highly correlated (at 0.99) within the range of the data. The marginal willingness to pay for housing computed from the regression procedure is also highly correlated with the true willingness to pay ($r=0.99$), and the mean value of the estimated willingness to pay is very close to the true value (the ratio is 0.99), at least for the 100 observations in the sample.

Now consider the quantal choice approach applied to the same data. For each of 100 households of differing income, the quantity of housing chosen and its price are both known. For each household, we randomly select four rejected dwellings and estimate the parameters of eq. (19) by maximum likelihood. Again, the average value of the estimated utility function is highly correlated with the known true value ($r=0.96$). The correlation of the computed with the actual marginal willingness to pay is somewhat lower ($r=0.90$), but the average values are close (the ratio is 0.99).

Tables 1 and 2 provide a summary comparison of the two methods of estimating willingness to pay. These measures are estimated for a single uniform income distribution (with y varying from 1 to 11 in units of 0.1), for four different housing distributions with mean 5, but with standard deviation of 2, 2.5, 4, and 8, and for one housing distribution with mean 11 and standard deviation 5. A 'housing market' is defined by drawing one hundred values of h from the distribution. Since the income distribution is the same for each housing market, those with relatively less variation in h are those where, ceteris paribus, the slope of the hedonic function is greater. In each case, the parameters of the utility function, the houses, and the income of occupants are sufficient to determine the equilibrium structure of housing prices. Willingness to pay is estimated from the set of 100 observations on income, housing, and housing prices.

Table 1 summarizes the estimates when the true utility function is GCES for a number of values of the underlying preference parameters. Panel A presents the correlations between the true marginal willingness to pay, $(\partial u/\partial h)/(\partial u/\partial y)$, and that estimated by the hedonic and quantal choice procedures. For 30 of 34 estimations by the hedonic method, the correlations of willingness to pay are almost exact ($r=0.99$). For three replications the correlations are close ($r=0.97$) and in one instance the correlation is far off

Table 1
Comparison of welfare measures for GCES utility functions estimated by hedonic and quantal choice methods.
(Mapping: $y = F(\mu, \sigma)$, where F is the cumulative normal density function.)

			Hedonic method/quantal choice method				
α	β	ε	$\mu=5, \sigma=2$	$\mu=5, \sigma=2.5$	$\mu=5, \sigma=4$	$\mu=5, \sigma=8$	$\mu=11, \sigma=5$
A. Correlations of true marginal willingness to pay with estimated values							
0.25	0.25	0.25	0.99/0.90	0.99/0.59	0.99/0.06	0.99/0.99	0.99/0.91
0.25	0.25	0.75	0.99/0.93	0.99/0.90	0.99/0.76	0.99/0.07	0.99/0.46
0.25	0.75	0.25	0.99/−0.02	0.99/0.18	0.99/0.89	0.99/0.99	0.99/0.75
0.25	0.75	0.75	0.97/0.07	0.99/0.79	0.99/0.98	0.99/0.98	0.97/0.05
0.75	0.25	0.25	0.99/0.04	0.99/0.64	0.99/0.95	0.99/0.99	0.99/0.83
0.75	0.25	0.75	0.99/0.01	0.99/0.07	0.99/0.96	0.99/0.99	0.72/0.36
0.75	0.75	0.25	0.99/0.15	0.99/0.64	0.97/0.69	0.99/0.93	–
B. Mean values of estimated marginal willingness to pay releative to true mean							
0.25	0.25	0.25	1.00/1.00	1.00/1.01	1.00/1.23	1.00/1.00	1.00/1.01
0.25	0.25	0.75	1.00/1.01	1.00/0.97	1.00/0.87	1.00/0.60	1.00/0.39
0.25	0.75	0.25	0.99/0.49	0.99/1.33	1.00/0.98	1.01/0.99	1.00/1.06
0.25	0.75	0.75	0.99/0.38	1.00/1.03	1.00/1.00	1.00/1.02	1.00/1.10
0.75	0.25	0.25	0.99/1.65	1.00/1.04	1.00/1.00	1.01/1.00	1.01/1.07
0.75	0.25	0.75	1.00/1.16	1.00/1.02	1.00/1.02	1.00/1.00	1.01/1.14
0.75	0.75	0.25	0.95/2.87	0.98/1.34	1.02/1.04	0.97/1.05	–

($r = 0.72$). In contrast, the correlations of the marginal willingness to pay estimated from the quantal choice model with the true values are often bizarre. For 13 of the 34 comparisons, the correlations are above 0.9, but for 10 of the comparisons it is below 0.1; in one case it is actually negative. Panel B compares the average values of the marginal willingness to pay for the different samples and estimating techniques. In all cases the mean value estimated by the hedonic technique is quite close to the true mean. It is never off by more than five percent. Again, the results for the quantal choice method are much more varied. In 21 of 34 cases, the estimated mean is within 10 percent of the true mean. In other cases, the mean is quite far off indeed. In one case, the estimated value is only 38 percent of the true value; in one case it is 287 percent. There is no pattern of deviation.

Table 2 provides a similar comparison of estimates when the underlying utility function is linear in paramaters. Results are presented for eight values of the underlying taste parameters for the same five mappings. Again comparisons are presented of the correlation of estimated and true marginal willingness to pay in each sample, and the relationships between the true mean willingness to pay and the estimated value. Despite the fact that the form of the utility function is linear, the estimates obtained by the hedonic method (which assumes they are GCES) are quite close. In 37 of the 40 replications, the correlations are 0.94 or better, and in two of the other cases the correlations are reasonable (i.e., $r = 0.80$, $r = 0.89$). In one case, the

Table 2
Comparison of welfare measures for linear utility functions estimated by hedonic and quantal choice methods.
(Mapping: $y=F(\mu,\sigma)$, where F is the cumulative normal density function.)

			Hedonic method/quantal choice method				
γ_1	γ_2	γ_3	$\mu=5,\sigma=2$	$\mu=5,\sigma=2.5$	$\mu=5,\sigma=4$	$\mu=5,\sigma=8$	$\mu=11,\sigma=5$
A. Correlations of true marginal willingness to pay with estimated values							
2.00	2.00	−0.25	0.99/0.97	0.98/0.79	0.99/0.89	0.99/0.99	0.99/0.96
2.00	2.00	−0.75	0.99/0.05	0.96/0.32	0.44/−0.13	0.97/0.92	0.90/0.15
2.00	4.00	−0.25	0.99/0.85	0.99/0.45	0.99/0.69	0.99/0.05	0.99/0.92
2.00	4.00	−0.75	0.99/0.90	0.99/0.60	0.99/0.06	0.99/0.99	0.99/0.95
4.00	2.00	−0.25	0.99/−0.07	0.98/−0.07	0.99/0.95	0.99/0.99	0.99/0.12
4.00	2.00	−0.75	0.99/0.11	0.94/−0.44	0.80/0.69	0.99/0.94	0.89/0.54
4.00	4.00	−0.25	0.99/0.93	0.99/0.61	0.99/0.05	0.99/0.99	0.99/0.99
4.00	4.00	−0.75	0.99/0.99	0.98/0.97	0.99/0.98	0.99/0.99	0.99/0.96
B. Mean values of estimated marginal willingness to pay relative to true mean							
2.00	4.00	−0.25	0.99/0.94	1.00/0.97	1.00/0.95	1.00/0.95	1.00/0.99
2.00	2.00	−0.75	0.99/0.01	1.00/1.14	1.01/0.92	1.01/0.87	1.01/1.35
2.00	4.00	−0.25	1.00/1.02	1.00/1.01	1.00/0.85	1.00/1.04	1.00/1.00
2.00	4.00	−0.75	1.00/0.97	1.00/0.98	1.00/0.87	1.00/0.91	1.00/0.95
4.00	2.00	−0.25	0.99/2.43	0.99/3.13	1.00/0.99	1.01/0.96	1.00/1.06
4.00	2.00	−0.75	0.98/0.31	0.99/1.41	1.01/0.97	1.01/0.92	1.01/1.39
4.00	4.00	−0.25	0.99/0.98	1.00/0.99	1.00/0.84	1.01/0.99	1.00/0.99
4.00	4.00	−0.75	0.99/1.05	1.00/0.99	1.00/0.94	1.00/0.92	1.00/0.94

correlation is very low, $r=0.44$. In contrast, the estimates obtained from the quantal choice function are again unexpected. In 18 of the 40 comparisons, the quantal choice method produces marginal willingness to pay estimates that are correlated at 0.9 or higher with the true values. In another three or four cases the correlations are reasonable, but in 10 cases the simple correlations are below 0.15. In four cases, the correlations are actually negative.

Similarly, comparisons of the mean values of the marginal willingness to pay, in Panel B, reveal that the hedonic method provides estimates reasonably close to the true average. In fact, in 40 comparisons only one is off by as much as 2 percent. In 27 cases the linear method results in an average willingness to pay within 10 percent of the true average. But in other cases it is quite far off – 141 percent, 243 percent, as much as 313 percent of the true mean.

5. Conclusion

It is obviously premature to draw firm conclusions from the few simulations presented in this paper. The numerical results so far, however, do not provide strong support for the robustness of the quantal choice technique when used to make welfare judgments about urban policies. In part, these results may have arisen because of the particular parameters or mapping used. The number of replications is rather small, especially by the standards of large Monte Carlo studies. In part, however, these results may merely indicate the fact well-known by macroeconomists: it is 'hard', statistically speaking, to estimate a function and to have any confidence in its rate of change.

In any case, the results suggest that extreme caution should be exercised in using these analytic techniques to make serious welfare comparisons.

References

Anas, Alex, 1980, Evaluating the effects of transportation–land use policies on housing values and household welfare, Environmental and Planning A 12, 747–764.

Anas, Alex, 1985, The capitalization of public investments: Transportation investment policy, in: John M. Quigley, ed., Perspectives on local public finance and public policy, Vol. 2 (JAI Press, Greenwich, CT) 59–80.

Burns, Leland S. and Leo Grebler, 1976, Resource allocation to housing investment: A comparative analysis, Economic Development and Cultural Change, Oct.

Case, Karl E., 1981, A new approach to modeling the effects of demographic change on the housing market, Working paper 42 (Department of Economics, Wellesley College, Wellesley, MA) March.

Ellickson, Bryan, 1981, An alternative test of the hedonic theory of housing markets, Journal of Urban Economics 9, no. 1, 56–79.

Harrison, David and Daniel L. Rubinfeld, 1978, Hedonic house prices and the demand for clean air, Journal of Environmental Economics and Management 5.

Kain, John F. and William C. Apgar, Jr., 1977, The modeling of neighborhood change, Discussion paper D77-22 (Department of City and Regional Planning, Harvard University, Cambridge, MA) Oct.

Kaufmann, Daniel and John M. Quigley, 1984, The consumption benefits of investment in urban infrastructure: The evaluation of sites and services programs in underdeveloped countries, Working paper 84-58, (Center for Real Estate and Urban Economics, University of California, Berkeley, CA).

Lerman, Stephen R., 1977, Location, housing, automobile ownership, and the mode to work: A joint choice model, Transportation Research Board Record, No. 610, 6–11.

McFadden, Daniel, 1974, Conditional logit analysis of qualitative choice behavior, in: Paul Zarembka, ed., Frontiers in econometrics (Academic Press, New York) 105–140.

McFadden, Daniel, 1978, Modelling the choice of residential location, in: Anders Karlqvist et al., eds., Spatial interaction theory and planning models (North-Holland, Amsterdam) 75–96.

Quigley, John M., 1976, Housing demand in the short run: An analysis of polytomous choice, Explorations in Economic Research 3, no. 1, 76–102.

Quigley, John M., 1982, Non linear budget constraints and consumer demand: An application to public programs for residential housing, Journal of Urban Economics 12, Oct., 177–201.

Quigley, John M., 1983, Estimates of a more general model of housing choice, in: Ronald E. Grieson, ed., The urban economy and housing (Lexington Books, Lexington, MA) 125–141.

Quigley, John M., 1985, Consumer choice of dwelling, neighborhood, and public services, Regional Science and Urban Economics 15, no. 1, 41–63.

Rosen, Sherwin, 1974, Hedonic prices and implicit markets: Product differentiation in pure competition, Journal of Political Economy 82, 34–55.

Williams, Jr., Roberton C., 1979, A logit model of demand for neighborhood, in: David Segal, ed., The economics of neighborhood (Academic Press, New York) 17–42.

Witte, A.D. et al., 1979, An estimate of a structural hedonic price model of the housing market: An application of Rosen's theory of implicit markets, Econometrica, Sept., 1151–1173.

Regional Science and Urban Economics 16 (1986) 43–56. North-Holland

AGGLOMERATIVE TENDENCIES IN THE DISTRIBUTION OF POPULATIONS

G. LEONARDI and J. CASTI

International Institute for Applied Systems Analysis, Laxenburg, A-2361, Austria

Received August 1984, final version received May 1985

The paper explores the role of accessibility to mutual contacts as an agglomeration force in the spatial distribution of population. The uniqueness conditions for the equilibrium solutions are analyzed in the static case, with the aid of mathematical programming embedding properties. A dynamic version in continuous time is then built, and conditions for instability of a globally stable equilibrium and appearance of multiple locally stable and unstable equilibria are stated. Finally, some implications for the geographical structure are discussed.

1. Introduction

In some recent work, Andersson and Ferraro (1982) and Ferraro (1984) have proposed and tested for Stockholm (with good empirical fit) a simple model for the equilibrium distribution of population in an urban area. The general idea behind the model, which departs a little bit from classic new urban economics, is that the city is essentially viewed as a public good [Shaked (1982)]. The public good feature is due essentially to spatial externalities which take the form of interactions among individuals. The propensity of people to interact with others constitutes an agglomerative force [Papageorgiou (1979)], which becomes the main factor shaping the landscape of the city.

This assumption is introduced in the model by making the residential attractiveness of a specific location within the city a function of both the dwelling space available (the private good component) and the accessibility to contact with others in different locations of the area (the public good component). More specifically, the model of Andersson and Ferraro takes the form

$$\left(\frac{Q_i}{P_i}\right)^{\alpha_1}\left(\sum_{j=1}^{n} f_{ij}P_j\right)^{\alpha_2}=K, \quad \text{where} \tag{1}$$

P_i = the relative population size in zone i, $P_i \geqq 0$, $\sum_i P_i = 1$,

Q_i = the dwelling space available in zone i,
f_{ij} = the spatial 'discount' factor associated with a move from zone i to zone j, $0 \leqq f_{ij} \leqq 1$,
α_1, α_2 = elasticity parameters,
K = a non-negative constant, $i = 1, 2, \ldots, n$.

The economic interpretation of (1) is as follows: the left-hand side of (1) represents the utility associated with choosing housing in zone j, which is dependent upon both the space per person available and the accessibility for contact with others. Eq. (1) states that at equilibrium, the utility level is spatially uniform and equals the constant level K in each zone.

With a small amount of algebra (1) can be rewritten in the form

$$P_i = Q_i\left(\sum_j f_{ij}P_j\right)^{\alpha} \Big/ \sum_k Q_k\left(\sum_j f_{kj}P_j\right)^{\alpha}, \qquad i = 1, 2, \ldots, n, \tag{2}$$

where $\alpha = \alpha_2/\alpha_1$. Eq. (2) defines a family of maps of a convex set D to itself,

$$F_\alpha : D \to D, \quad \text{where}$$

$$D = \left\{ P \in R^n : P_i \geqq 0, \sum_i P_i = 1 \right\}.$$

In the Stockholm application, the above mentioned authors do not solve eq. (2), but a linear approximation to it, in order to guarantee uniqueness of equilibrium. However, we feel that this simplification neglects the most interesting part of (2), namely the possibility of having multiple equilibria.

Therefore, our objectives in this paper are (i) to study the fixed points of the map F_α as a function of the elasticity ratio α; (ii) to provide an alternate interpretation of (2) viewing it as a mathematical programming problem; (iii) to provide a dynamical framework, based upon a generalized epidemic model, whose equilibrium behavior coincides with (2).[1]

The primary reason for undertaking the following analysis is to gain a deeper understanding into the appearance of multiple equilibria (more than one fixed point) for (2) as the elasticities vary and to examine the effect changes in the elasticities have upon the dynamic version of (1) constructed below. Our basic conclusion is that there exists a physically realizable set of elasticities α_1^*, α_2^* such that below the ratio $\alpha^* = \alpha_2^*/\alpha_1^*$ any initial population distribution converges to a single equilibrium distribution, while above α^*

[1]As pointed out by an anonymous referee, the solution to an equation which resembles (2) has been studied in Beckmann (1976); dynamic aspects of a related formulation may also be found in Harris and Wilson (1978).

many different equilibria may emerge, depending upon the particular α and the initial distribution.

2. A rational expectations interpretation

Before investigating the fixed point behavior of (2) it is of considerable interest to consider an alternate interpretation of the basic model of Andersson and Ferraro, an interpretation based upon expectations over future moves in the urban area, rather than the Andersson–Ferraro interpretation based upon the notion of accessibility.

Suppose an individual settles in zone i by maximizing his utility subject to a random-utility type of heterogeneity. The utility of a choice to settle in zone i is evaluated taking into account *all* future moves that could be made from zone i to any other zone. Assume the time horizon is N periods, where N is fixed and finite. Let

$V_i(N)$ = expected utility of choosing zone i when N decisions remain.

Then an elementary application of Bellman's Principle of Optimality yields the relation

$$V_i(N) = \mathop{\mathrm{E}}_{\varepsilon}\left\{\max_{(j)}\left[-c_{ij}+\varepsilon_{ij}+\alpha V_j(N-1)\right]\right\}, \quad \text{where}$$

c_{ij} = the cost of moving from zone i to zone j,
ε_{ij} = the random utility associated with a move from zone i to zone j,
α = a discount factor on the future ($\alpha<1$),
E_{ε} = expectation over ε_{ij},
$\max_{(j)}$ = maximization over *all alternatives* in zone j, and over *all zones* j.

Assume now that the number of alternatives available in zone j is Q_j, where Q_j is 'large'. Then using extreme value asymptotic estimates [Leonardi (1982, 1983a)], the equation in $V_i(N)$ takes the form

$$V_i(N) = \log \sum_j f_{ij} Q_j \exp(\alpha V_j(N-1)). \tag{3}$$

[For the method of deriving recurrence equations like (3), see Leonardi (1983b)]. Now let $N \to \infty$ and set $x_i = \lim_{N\to\infty} V_i(N)$ (if it exists). Then we have

$$x_i = \log \sum_j f_{ij} Q_j \exp(\alpha x_j). \tag{4}$$

But, if we set

$$\phi_j = \sum_j f_{ij} P_j, \qquad v_i = \log \phi_i, \qquad A = \sum_j Q_j \phi_j^\alpha$$

in the Andersson–Ferraro model (1), then (1) becomes

$$v_i = \log \sum_j f_{ij} Q_j \exp(\alpha v_j) - \log A,$$

which is the same as (4) if we neglect the irrelevant constant term $\log A$.

Once the x_i are obtained from (4), suppose a pool of potential settlers from the rest of the world are deciding where to settle in the region. If the cost of moving from anywhere in the world to any zone i in the region is c, then the utility u_i of such a choice is

$$u_i = -c + \varepsilon_i + \alpha x_i,$$

where ε_i is again a random utility term. If we maximize this utility over all alternatives, we obtain the logit formula

$$P_i = Q_i \phi_i^\alpha \Big/ \sum_j Q_j \phi_j^\alpha.$$

This completes the proof of the equivalence of the Andersson–Ferraro interpretation with the above rational expectations interpretation.

We can now give a mathematical programming reformulation of the basic problem. If we consider eq. (3) as the *dual* program to a certain *primal* mathematical programming problem, we can construct the primal as the dynamic program

$$V_N(P) = \max_{q_{ij}} [H(P, q_{ij}) + \alpha V_{N-1}(R)], \quad \text{where} \tag{5}$$

$$R_j = \sum_i P_i q_{ij}, \qquad \sum_j q_{ij} = 1, \qquad 0 < \alpha < 1, \qquad V_0(P) = 0.$$

The sequence $\{V_N(P)\}$ is related to the sequence $\{V_i(N)\}$ as

$$V_N(P) = \sum_i P_i V_i(N),$$

where the function $H(P, q_{ij})$ is given by

$$H(P, q_{ij}) = -\sum_i P_i \sum_i q_{ij} \log(q_{ij}/f_{ij} Q_j).$$

Thus, the programs (3) and (5) are precise duals.

We now inquire as to what program has eq. (1) as the first-order necessary conditions for its optimal solution. The answer is

Proposition 1. The first-order necessary conditions (in the P variables) for the problem

$$\max_{P,D,q}\left\{-\sum_i P_i \log P_i/Q_i - \sum_i D_i \sum_j q_{ij} \log q_{ij}/f_{ij}Q_j\right\},$$

when $D_j = \alpha(P_j + \sum_i D_i q_{ij})$, $\sum_i q_{ij} = 1$, $0 < \alpha < 1$, *are equivalent to eq. (2) for P.*

Unfortunately, in general the objective function for this program is not concave and the program is only defined for $0 < \alpha < 1$, both factors seriously limiting the utility of the program in the study of eq. (2). Now let us return to the analysis of the fixed points of the map F_α generated by eq. (2).

3. Existence and uniqueness of equilibria

As discussed in section 2, eq. (2) for the equilibrium population distribution is equivalent to eq. (4). Thus, we study the properties of the maps

$$G_\alpha : R^n \to R^n,$$

given by

$$(G_\alpha)_i(x) = \log \sum_j f_{ij} Q_j \exp(\alpha x_j), \qquad i = 1, 2, \ldots, n.$$

Our first result is

Proposition 2. There exists a unique solution to the equation

$$G_\alpha(x) = x$$

for each α *in the range* $0 \leqq \alpha < 1$.

Proof. By the Mean-Value Theorem, for $x, y \in R^n$, there exists an $\xi \in R^n$ on the line joining x and y such that

$$G_i(y) - G_i(x) = \sum_j G_{ij}(y_j - x_j),$$

where $G_{ij} \doteq \partial G_i / \partial \xi_j$. Using the norm

$$\|v\| = \max_j |v_j|, \qquad v \in R^n,$$

we have

$$\sum_j G_{ij}(y_j - x_j) \leqq \|y-x\| \cdot \sum_j G_{ij}.$$

From which it follows that

$$\|G(y)-G(x)\| \leqq \max_i \left|\sum_j G_{ij}\right| \|y-x\|.$$

But, we also have

$$G_{ij} = \alpha f_{ij} Q_j \exp(\alpha \xi_j) \Big/ \sum_j f_{ij} Q_j \exp(\alpha \xi_j)$$

and $\sum_j G_{ij} = \alpha$. Hence,

$$\|G(y)-G(x)\| \leqq \alpha \|y-x\|,$$

implying that G is a contraction for $0 \leqq \alpha < 1$, completing the proof by an appeal to the Contraction Mapping Theorem. □

The problem with the above result is that it is valid for too small a range of elasticities α. We now inquire as to when there is still a unique equilibrium if $\alpha \geqq 1$. To address this question, we need the following *global* univalence theorem.[2]

Global Uniqueness Theorem. Let f be a continuously differentiable map of a bounded convex open set $D \subset R^n$ into R^n. Then f is globally one-to-one if all principal minors of the symmetric part of $J = (\partial f / \partial x)$ are positive in D.

[Note: the symmetric part of J is $1/2(J + J')$].

Proof. It is only necessary to show that if $f(x) = f(y)$ for $x, y \in D$, then $x = y$. This is certainly true if $f(x) = \text{grad}\, F(x)$ and $F(x)$ is strictly convex, since if $F(x)$ is strictly convex and $x \neq y$,

$$F(x) + f(x) \cdot (y - x) < F(y),$$

and $f(x) = f(y)$ would imply $x = y$. □

[2]To economists, the best known presentation of univalence theorems is found in Nikaido (1968).

A necessary and sufficient condition that $F(x)$ be strictly convex in D is that the Hessian matrix

$$H(x)=\frac{\partial^2 F}{\partial x^2}>0.$$

But, by a result of Frobenius, a real symmetric matrix A is positive definite if, and only if, all principal minors of A are positive. Hence, $f(x)$ will be globally one-to-one if all principal minors of $F(x)$ are positive when $f(x)$ is a gradient system.

To extend this result to non-gradient systems $f(x)$, suppose $x, y \in D$ and set

$$q(t)=tx+(1-t)y,$$

$$\phi(t)=[f(q(t))-f(y)]\cdot(x-y), \qquad 0\leqq t\leqq 1.$$

Since D is convex, $q(t)\in D$. We have $\phi(0)=0$ and $\phi(1)\neq 0$ implies $f(x)\neq f(y)$. Upon differentiating $\phi(t)$, we obtain

$$\phi'(t)=\sum_{i,j}\frac{\partial f_i(q(t))}{\partial x_j}(x_i-y_i)(x_j-y_j)\geqq K|x-y|^2>0.$$

Thus, $\phi'(t)>0$ for $0<t\leqq 1$. By the Mean Value Theorem

$$\phi(1)=\phi(0)+\phi'(\theta)>0, \qquad 0<\theta<1.$$

Thus, $\phi(1)>0$ and $f(x)\neq f(y)$, which implies f is one-to-one in D.

We consider application of this result to the system (2), which can be written as

$$F_i(P)=Q_i\phi_i^\alpha-P_i\sum_j Q_j\phi_j^\alpha=0, \quad \text{where}$$

$$\phi_i=\sum_j f_{ij}P_j.$$

The Jacobian matrix at the equilibrium P^* is

$$\left(\frac{\partial F_i}{\partial P_j}\right)(P^*)=A\left[\alpha P_i/P_j\left(q_{ij}-\sum_k q_{kj}P_k\right)-\delta_{ij}\right], \quad \text{where}$$

$$A=\sum_j Q_j\phi_j^\alpha, \qquad q_{ij}=\frac{f_{ij}P_j}{\sum_j F_{ij}P_j}, \qquad \delta_{ij}=1 \text{ if } i=j, \quad =0 \text{ if } i\neq j.$$

If we denote the symmetric part of J as $\hat{J}$, then for a 2-zone case we have

$$\hat{J}_{11}(P^*)=A\left[\alpha\left(q_{11}-\sum_k q_{k1}P_k\right)-1\right]$$

and the condition $\hat{J}_{11}>0$ implies

$$\alpha>1\bigg/\left(q_{11}-\sum_k q_{k1}P_k\right).$$

The other condition for uniqueness in the 2-zone case is $\det\hat{J}>0$. We have

$$\det\hat{J}=\frac{A^2}{4}\left[2\alpha\left(q_{11}-\sum_k q_{k1}P_k\right)-1\right]\left[2\alpha\left(q_{21}-\sum_k q_{k2}P_k\right)-1\right]$$
$$-\frac{A^2}{4}\left[\alpha(P^1/P_2+P_2/P_1)(\tfrac{1}{2}q_{12}+\tfrac{1}{2}q_{21})-\tfrac{1}{2}\sum_k q_{k2}P_k-\tfrac{1}{2}\sum_k q_{k1}P_k\right]^2.$$

The requirement $\det\hat{J}>0$ leads to the condition on α:

$$a\alpha^2+b\alpha+c>0,\quad \text{where}$$

$$a=(q_{11}-\sum\varepsilon q_{k1}P_k)\left(q_{22}-\sum_k q_{k2}P_k\right)-\tfrac{1}{4}(p_1/p_2+P_2/P_1)^2\cdot(q_{12}+q_{21})^2$$

$$b=-\tfrac{1}{2}(P_1/P_2+P_2/P_1)(q_{12}+q_{21})\left(\sum_k q_{k2}P_k+\sum_k q_{k1}P_k\right)$$
$$-2\left(q_{11}-\sum_k q_{k1}P_k\right)-2\left(q_{22}-\sum_k q_{k2}P_k\right)$$

$$c=l-\tfrac{1}{4}(\sum\varepsilon_{k1}P_k+\sum q_{k2}P_k)^2.$$

It is clear that for an n-zone case, there will be n polynomial inequalities in α of the fpregoing type that would suffice to insure a globally unique equilibrium. Since the algebra involved in the general case is quite imposing, we shall not prove it here but leave it for a suitable computer symbolic manipulation program.

4. A dynamic generalization

Thus far, we have considered only the behavior of the solutions of the *static* model (2). In this section we derive a dynamic model, based upon epidemiological considerations, whose equilibrium behavior coincides with (2). While there may be an infinite number of dynamic models possessing this same long-run behavior, we feel that the natural abstract similarities between the way populations spread out in space and the way infectious diseases move through a spatially-distributed population are strong enough to provide considerable support for this model.

We begin with the general population dynamics,

$$\frac{\mathrm{d}P_j}{\mathrm{d}t} = \sum_i P_i r_{ij} - P_j \sum_i r_{ji}, \qquad i, j = 1, 2, \dots, n,$$

expressing the simple fact that the rate of change of populations in zone i is just the rate of immigration minus the rate of emigration. The quantities r_{ij} represent the fraction of the population in zone i who want to move to zone j. If we choose the r_{ij} to imitate an epidemiological interaction, then we have

$$r_{ij} = \lambda Q_j \phi_j^\alpha,$$

where again $\phi_i = \sum_j f_{ij} P_j$. Roughly speaking, this interaction just says that the rate of immigration is proportional both to the housing available and the accessibility to contacts with others, with λ being a proportionality constant. With foregoing choice of r_{ij}, we are led to the dynamics

$$\frac{\mathrm{d}P_j}{\mathrm{d}t} = Q_j \phi_j^\alpha - P_j \sum_i Q_i \phi_i^\alpha,$$

where, for simplicity, we set $\lambda = 1$.

Another way of arriving at the same result is to argue as follows. Let there be j alternatives to be adopted by a given population and let $P_j(t)$ be the number of people adopting alternative j at time t. To begin with, suppose an individual currently using alternative i can switch to j in a small time interval $(t, t+\Delta)$ with probability $\lambda \Delta h_{ij} Q_j P_j(t)$, where

λ = speed parameter,
h_{ij} = a decreasing function of the cost of switching from i to j, $h_{ij} = h_{ji}$, $h_{ij} \geqq 0$,
Q_j = the 'capacity' of alternative j.

We are then led to the dynamic equations

$$\frac{\mathrm{d}P_j}{\mathrm{d}t} = \lambda \sum_i P_i(t) h_{ij} Q_j P_j(t) - \lambda P_j(t) \sum_i h_{ji} Q_i P_i(t).$$

As a second step, let us assume that the probability of switching from i to j is proportional not to the number of adopters of j, but to the 'accessibility' from j to the adopters of all other alternatives. Thus, instead of the above probability of switching, we get

$$\lambda \Delta h_{ij} Q_j \left(\sum_k f_{jk} P_k(t) \right)^{\alpha},$$

where f_{ij} is as before. The corresponding dynamics now are

$$\frac{\mathrm{d}P_j}{\mathrm{d}t} = \lambda \sum_i P_i(t) h_{ij} Q_j \phi_j^{\alpha} - \lambda P_j(t) \sum_i h_{ji} Q_i \phi_i^{\alpha}, \tag{6}$$

where $\phi_i(t) = \sum_j f_{ij} P_j(t)$. The dynamics (6) can easily be seen to have (2) as its equilibrium behavior. Now let us examine the stability behavior of an equilibrium for (6).

We consider the simplified dynamics where $h_{ij} = 1$, $\lambda = 1$,

$$\dot{P}_j = Q_j \phi_i^{\alpha} - P_j \sum_i Q_i \phi_i^{\alpha}.$$

Computing the system Jacobian matrix F at the equilibrium $\dot{P} = 0$, after some algebra we obtain

$$[F]_{ij} = A \left\{ \alpha \frac{P_i}{P_j} (q_{ij} - \sum_k P_k q_{kj}) - \delta_{ij} \right\}, \quad \text{where}$$

$$q_{kj} = f_{kj} P_j \Big/ \sum_m f_{km} P_m \quad \text{and} \quad A = \sum_i Q_i \phi_i^{\alpha}.$$

Define the new matrices B and C as

$$[B]_{jk} = \frac{P_j}{P_k} \left(q_{jk} - \sum_i P_i q_{ik} \right),$$

$$C = \alpha B - I.$$

It is clear that the characteristic values of F are related to those of C as

$$\lambda_F = A \lambda_C \quad \text{and} \quad \lambda_C = \alpha \lambda_B - 1.$$

Thus, the stability condition $\operatorname{Re} \lambda_F < 0$ translates into

$$\operatorname{Re}\lambda_B < 1/\alpha \quad \text{or} \quad \alpha < \frac{1}{\sup \operatorname{Re}\lambda_B}.$$

However, with a little algebra, it can be shown that one root of B is 0, while the remaining $n-1$ roots are the same as those of the matrix

$$[Q]_{ij} = q_{ij}$$

as defined above, omitting the root $\lambda_Q = 1$. Note that, since Q is a stochastic matrix, one root is real and equal to 1, while $|\lambda_Q| < 1$ for the remaining $n-1$ roots.

Thus, we can summarize the stability by saying that there exists a value $\bar{\alpha}$ of α such that

$$\bar{\alpha} = \frac{1}{\sup \lambda(Q)}, \qquad 1 \leqq \bar{\alpha} \leqq \infty,$$

and that for all $\alpha \leqq \bar{\alpha}$ there is a unique globally stable equilibrium, while for $\alpha > \bar{\alpha}$ there are multiple equilibria, each of which may be locally stable or unstable. Furthermore, if we take special cases of $\{f_{ij}\}$, then we obtain the sharper results

$$f_{ij} = \delta_{ij} \Rightarrow \bar{\alpha} = 1 \qquad \text{(strong locational friction)},$$

$$f_{ij} = 1 \Rightarrow \bar{\alpha} = \infty \qquad \text{(no friction)},$$

$$f_{ij} = \text{general} \Rightarrow 1 < \bar{\alpha} < \infty \qquad \text{(smooth friction)}.$$

5. Concluding remarks

Suggestions for further research, rather than concluding remarks, are perhaps the main results of the exploratory analysis carried out in this paper. The main question addressed in the beginning, concerning conditions for existence of single rather than multiple solutions, has been partially answered. The mathematical programming formulation has provided indirectly a sufficient condition for global uniqueness (the elasticity ratio being less than one), while the dynamic analysis has shown that a threshold for the elasticity ratio exists, which is in general greater than one, separating the region of a unique stable solution from that of multiple locally stable and unstable solutions.

From the point of view of the insight into the consequences on the geographical structure, however, still much remains to be done. An interesting result is the dependence of the threshold upon the spatial frictions

f_{ij}, measuring the deterrence effect on interaction of spatial separation. The main conclusion is that the smaller this effect, the wider the range of values for the elasticity ratios yielding a unique equilibrium. It is not clear yet, however, what the shape of the corresponding population distribution is below and beyond the threshold. Intuitively, the nature of the problem seems to suggest that the unique equilibrium below the threshold corresponds to a 'dispersed' pattern (as dispersed as can be in a bounded region), with densities decreasing smoothly from a single 'center' (the most accessible zone) to the periphery. On the other hand, beyond the threshold the 'dispersed' solution becomes unstable and other locally stable solutions eventually appear, with peaks in zones other than the 'central' one, and which pattern will prevail depends on the initial conditions. In order to make this conjecture more precise, if numerical simulation is to be avoided, one has to introduce sharper assumptions on the topology of the geographic space and the structure of the f_{ij}. It is interesting to note that such an approach, for a problem closely related to ours, has been developed in Papageorgiou and Smith (1983). It is also surprising to note that the above two authors seem to arrive at conclusions opposite to ours, namely, the smaller the distance deterrence effect, the smaller the region of stability for a dispersed equilibrium. But this might of course be due to differences in the functional specifications of the two models.

In any event, a partial confirmation of our conjectures on the spatial pattern is provided by analyzing the simple case of two zones of identical size. The calculations are straightforward and will not be given here. Only a summary of the results in graphical form will be shown. The threshold parameter $\bar{\alpha}$ is found to be

$$\bar{\alpha}=(1+f)/(1-f),$$

where f is the distance discount factor between the two zones. Clearly,

$$\bar{\alpha}=1 \quad \text{for} \quad f=0,$$

$$\bar{\alpha}=\infty \quad \text{for} \quad f=1,$$

as required. The behavior of the rate of change of the relative population in one zone (the first, say) is shown in fig. 1.

It is thus seen that, for $\alpha \leqq \bar{\alpha}$, the uniform distribution $p=1/2$ is actually globally stable, corresponding to the graphs (a) and (b). For $\alpha > \bar{\alpha}$, the situation is depicted in graph (c). The uniform distribution becomes unstable, and two locally stable solutions $\bar{P}_1$ and $\bar{P}_2$ appear, with domain of attraction $0 \leqq P < \frac{1}{2}$ and $\frac{1}{2} < P \leqq 1$, respectively. That is, if the system starts with more

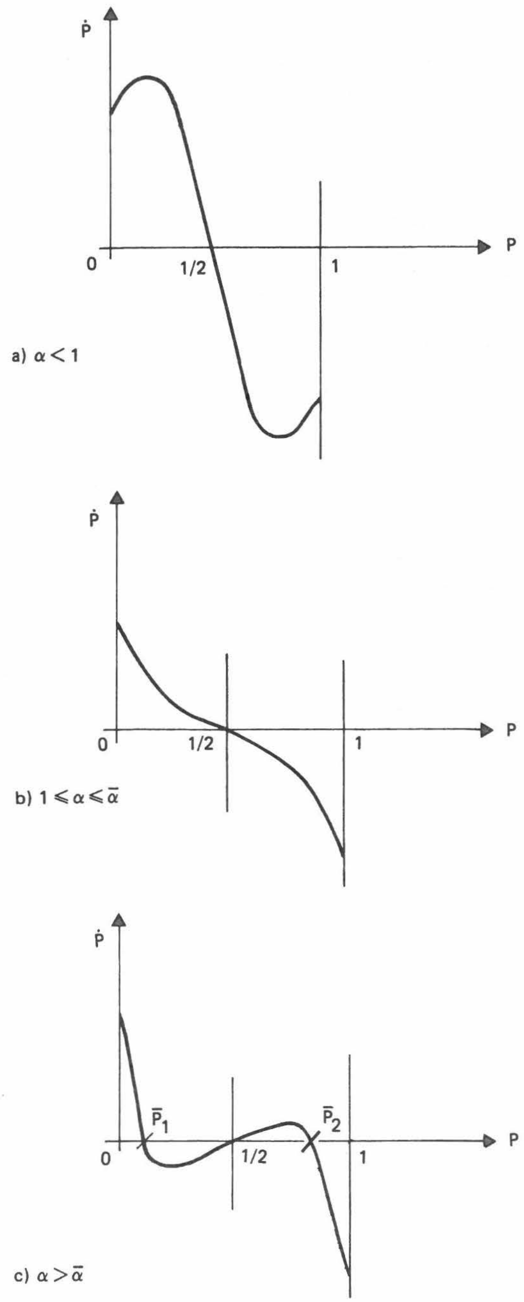


Fig. 1. Stability analysis for the two zones case.

population in one zone, that zone will become the center of the system, with the highest population density.

Among other things, it is surprising how such an asymmetric behavior arises from a purely symmetric environment.

References

Andersson, Å.E. and G. Ferraro, 1982, Accessibility and density distribution in metropolictan areas: Theory and empirical studies, Umeå Economic Papers CIX, forthcoming in: Papers of the Regional Science Association.

Beckmann, M.J., 1976, Spatial equilibrium in the dispersed city, in: G.J. Papageorgiou, ed., Mathematical land use theory (Lexington Books, Lexington, MA) 117–125.

Ferraro, G., 1984, Localizzazioni residenziali entro aree metropolitane: Comportamenti individuali et assetti di equilibrio (Residential Location in Metropolitan Areas: Individual Behavior and Equilibrium Pattern), Working paper (DAEST, Venice).

Harris, B. and A.G. Wilson, 1978, Equilibrium values and dynamics of attractiveness terms in production-constrained spatial-interaction models, Environment and Planning A 10, 371–388.

Leonardi, G., 1982, The structure of random utility models in the light of the asymptotic theory of extremes, WP-82-91 (International Institute for Applied Systems Analysis, Laxenburg, Austria).

Leonardi, G., 1983a, Transient and asymptotic behavior of a random-utility based stochastic search process in continuous space and time, WP-83-108 (International Institute for Applied Systems Analysis, Laxenburg, Austria).

Leonardi, G., 1983b, An optimal control representation of a stochastic multistage-multiactor choice process, in: D.A. Griffith and A.C. Lea, eds., Evolving geographical structures (Martinus Nijhoff, The Hague).

Nikaido, H., 1968, Convex structures and economic theory (Academic Press, New York).

Papageorgiou, Y.Y., 1979, Agglomeration, Regional Science and Urban Economics 9, no. 1, 41–59.

Papageorgiou, Y.Y. and T.R. Smith, 1983, Agglomeration as local instability of spatially uniform steady-states, Econometrica 51, no. 4, 1109–1119.

Shaked, A., 1982, Human environment as a local public good, Journal of Mathematical Economic Theory 9, 418–448.

Regional Science and Urban Economics 16 (1986) 57–80. North-Holland

A DYNAMIC MIGRATION THEORY AND ITS EVALUATION FOR CONCRETE SYSTEMS

G. HAAG and W. WEIDLICH

Universität Stuttgart, D-7000 Stuttgart 80, FRG

Received October 1984, final version received April 1985

This paper presents a theoretical framework for analyzing multiregional migration as a stochastic process. The equation of motion is formulated as a master equation. A quasi-deterministic meanvalue equation is derived from the master equation. The analysis is focused on the solution of the meanvalue equation. Finally it is described how the approach can be applied to empirical migration data in a study of migration processes in Canada for which the migratory stress is evaluated. The relation to static random utility theory is also established.

1. Introduction

Let us consider the migration of one homogeneous population consisting of N members between L areas. The quantity of particular interest is the 'migratory configuration'

$$\{n_1, n_2, \ldots, n_L\}, \qquad \sum_{l=1}^{L} n_i = N, \tag{1.1}$$

where n_j is the number of individuals in area j at a given time. The evolution of these numbers n_j is a stochastic, i.e., probabilistic process, since the migration of people depends on individual decisions.

The fully adequate treatment of this process is therefore one, which takes into account the random fluctuations of the n_j. This treatment starts with introducing the probability distribution

$$P(n_1, n_2, \ldots, n_L, t) \equiv P(\boldsymbol{n}, t), \tag{1.2}$$

which is by definition the probability to find the migratory configuration $\{n_1, \ldots, n_L\}$ at time t. The equation of motion for $P(\boldsymbol{n}, t)$ is noted as 'master equation' (section 1.1).

Below this stochastic level there exists a quasideterministic level of description restricted to an equation of motion for meanvalues. This meanvalue equation can be derived from the master equation (section 1.2). In this article we shall focus on the solution (section 1.4) and evaluation (section 2) of the meanvalue equation, which is sufficient in most cases. A first step towards the application of the evaluation methods to the Canadian migratory system is made in section 3.

1.1. The master equation for the stochastic evolution of the migratory configuration

The equation of motion for $P(\boldsymbol{n}, t)$ reads [for details, see Haag and Weidlich (1984)]

$$\frac{\mathrm{d}P(\boldsymbol{n},t)}{\mathrm{d}t}=\sum_{i,j=1}^{L} w_{ij}(\boldsymbol{n}^{(ji)})P(\boldsymbol{n}^{(ji)},t)-\sum_{i,j=1}^{L} w_{ji}(\boldsymbol{n})P(\boldsymbol{n},t), \tag{1.3}$$

with the following notation and interpretation:

$$\boldsymbol{n}^{(ji)}=\{n_1,n_2,\ldots,(n_j+1),\ldots,(n_i-1),\ldots,n_L\} \tag{1.4}$$

is a migratory configuration neighboring to

$$\mathbf{n}=\{n_1,n_2,\ldots,n_j,\ldots,n_i,\ldots,n_L\}.$$

$w_{ji}(\boldsymbol{n})$ is the transition probability per unit of time from configuration $\boldsymbol{n}$ to configuration $\boldsymbol{n}^{(ji)}$ via migration of any one of the n_i members in area i to area j. Since each of the n_i members in i has an independent individual transition probability per unit of time to migrate from i to j, namely p_{ji}, we find[1]

$$w_{ji}(\boldsymbol{n})=n_i\cdot p_{ji}. \tag{1.5}$$

The p_{ji} in general can also depend on the configuration $\boldsymbol{n}$.

The interpretation of (1.3) is straightforward. The probability $P(\boldsymbol{n},t)$ of $\boldsymbol{n}$ changes with time for two reasons: (1) because of the probability flux into $\boldsymbol{n}$ from neighboring configurations $\boldsymbol{n}^{(ji)}$ [first sum on the right-hand side of (1.3)] and (2) because of the probability flux from $\boldsymbol{n}$ into neighboring configurations $\boldsymbol{n}^{(ji)}$ [second sum on the right-hand side of (1.3)]. Both terms are counteractive and contribute to the evolution of $P(\boldsymbol{n},t)$. The equilibrium, i.e., $\mathrm{d}P_{\mathrm{st}}(\boldsymbol{n})/\mathrm{d}t=0$, is defined by the balance of in- and out-fluxes of the configuration $\boldsymbol{n}$.

[1]We prefer to write the index i of the origin area right of the index j of the destination area in the transition probabilities, because this is more convenient in matrix equations. This differs however from the conventional notation.

1.2. The equation of motion for the meanvalues of the migratory configuration

The master equation (1.3) yields an equation of motion for the meanvalues,

$$\bar{n}_i(t) = \sum_{\boldsymbol{n}} n_i P(\boldsymbol{n}, t), \tag{1.6}$$

which reads [for details of the derivation, see Haag and Weidlich (1984)]

$$\begin{aligned}\frac{\mathrm{d}\bar{n}_k(t)}{\mathrm{d}t} &= \sum_{j=1}^{L} p_{kj}\bar{n}_j(t) - \sum_{j=1}^{L} p_{jk}\bar{n}_k(t) \\ &= \sum_{j=1}^{L} w_{kj}(\bar{\boldsymbol{n}}) - \sum_{j=1}^{L} w_{jk}(\bar{\boldsymbol{n}}) \\ &\equiv W\ (\text{into } k) - W\ (\text{out of } k).\end{aligned} \tag{1.7}$$

The individual transition probabilities p_{kj} here can still depend on the mean configuration $\{\bar{n}_1, \ldots, \bar{n}_L\}$. Henceforth, however, we shall treat the case, where the p_{kj} are independent of $\bar{\boldsymbol{n}}$ and hence the eqs. (1.7) are linear in $\bar{n}_k$. The p_{kj} may, however, still depend on time by exogenous influences.

Since the terms with $j=k$ cancel on the right-hand side of (1.7), we can put $p_{kk} \equiv 0$ without changing the dynamics. From now on we skip the bar in $\bar{n}_k(t)$ writing $n_k(t)$ for the meanvalues.

The structure and interpretation of the m.v.e. (meanvalue equations) (1.7) is similar to that of the m.e. (master equation) (1.3) of the full stochastic process: evidently, $w_{kj} = p_{kj}n_j$ is the mean number of people migrating per unit of time from j to k; the corresponding holds for $w_{jk} = p_{jk}n_k$. The change with time of $n_k(t)$ is therefore, according to (1.7), due to the flux of people W (into k) minus the flux of people W (out of k).

If the p_{kj} are independent of $\boldsymbol{n}$, as presumed, the eqs. (1.7) have *again* the structure of a standard master equation.[2] This becomes clear if we introduce the *relative frequencies* $p(k, t)$ to find a member of the total population in area k,

$$p(k,t) = \frac{n_k(t)}{N}, \qquad \sum_{j=1}^{L} p(j,t) = 1. \tag{1.8}$$

For them, eq. (1.7) reads

$$\frac{\mathrm{d}p(k,t)}{\mathrm{d}t} = \sum_{j=1}^{L} p_{kj}p(j,t) - \sum_{j=1}^{L} p_{jk}p(k,t). \tag{1.9}$$

[2]If, on the other hand, p_{kj} depends on $\boldsymbol{n}$, the eqs. (1.7) become non-linear and are not equivalent to a master equation, which is always linear in the variable. The case $p_{kj}(\boldsymbol{n})$ is important for the treatment of agglomeration processes [see Weidlich and Haag (n.d.)].

The $p(k,t)$ evidently can also be interpreted as the *probability* to find an individual in area k at time t. Then (1.9) is a standard master equation for the $p(k,t)$, and will henceforth be considered in this form.

Let us now introduce a probability with a special initial condition, namely the *conditional probability* $p(k,t\,|\,i,0)$ with the following defining properties:

$$p(k,0\,|\,i,0)=\delta_{ki}, \tag{1.10a}$$

$$\sum_{k=1}^{L} p(k,t\,|\,i,0)=1, \tag{1.10b}$$

$$p(k,t\,|\,i,0) \text{ obeys eq. (1.9).} \tag{1.10c}$$

The conditional probability is the probability to find an individual in region k at time t, given that he/she was in region i at time $t=0$.

Making use of the conditional probability, a probability distribution $p(k,t)$ with *arbitrary* initial values $p(k,0)$ can easily be constructed:

$$p(k,t)=\sum_{i=1}^{L} p(k,t\,|\,i,0)p(i,0). \tag{1.11}$$

Furthermore, the conditional probability satisfies the constitutive Chapman–Kolmogorov equation,

$$p(k,t_3\,|\,i,t_1)=\sum_{j=1}^{L} p(k,t_3\,|\,j,t_2)p(j,t_2\,|\,i,t_1), \tag{1.12}$$

which says: the probability to be in k at time t_3, having been in i at time t_1 [left-hand side of (1.12)] can be composed of all probabilities to reach any intermediate state j at time t_2 from state i at time t_1 and, proceeding from j, subsequently to reach state k at time t_3. It can be shown, that (1.12) is a consequence of the master equation (1.9) and that vice versa the master equation (1.9) can be derived from (1.12).

The conditional probability $p(k,t\,|\,i,0)$ should not be confused with the transition probability per unit of time p_{ki}! This is already obvious for dimensional reasons: whereas the probability $p(k,t\,|\,i,0)$ is a dimensionless number, the transition probability p_{ki} is a probability per unit of time, or a probability rate with the dimension $[t^{-1}]$. To see the relation between both quantities, let us apply the master equation (1.9) to $p(k,t\,|\,i,0)$ for an infinitesimally small time interval $t \Rightarrow \Delta t$, writing

$$\frac{\mathrm{d}p(k,t\,|\,i,0)}{\mathrm{d}t} \Rightarrow \frac{p(k,\Delta t\,|\,i,0)-p(k,0\,|\,i,0)}{\Delta t}$$

on the left-hand side of (1.9). Solving for $p(k, \Delta t \mid i, 0)$ and inserting (1.10a, b) on the right-hand side of (1.9), one obtains

$$p(k, \Delta t \mid i, 0) = \Delta t \cdot p_{ki} \quad \text{for} \quad k \neq i, \tag{1.13a}$$

$$p(i, \Delta t \mid i, 0) = 1 - \Delta t \sum_{j=1}^{L} p_{ji}. \tag{1.13b}$$

Evidently, $p(k, \Delta t \mid i, 0)$, $k = 1, 2, \ldots, L$, fulfills the probability normalization condition (1.10b). On the other hand, we mention that the transition probabilities p_{ji} *do not have to fulfill* any probability normalization conditions (apart from being positive definite quantities).

From (1.13a, b) it is clear, that $p(k, \Delta t \mid i, 0)$ is proportional to p_{ki} (for $k \neq i$) for *infinitesimal* times Δt only, whereas the composition of the conditional probability $p(k, t \mid i, 0)$ in terms of the transition probabilities p_{ki} for *finite* times t is much more complicated and is determined by the equation of motion (see section 1.4).

For *general* transition probabilities only few analytic statements about the solutions $p(k, t)$ of (1.9) are possible. Most important are the following results:

(1) If all p_{ki} are non-vanishing and constant with time, there exists a unique stationary solution $p_{st}(k)$ of the master equation.
(2) Any time dependent solution $p(k, t)$ approaches $p_{st}(k)$ for $t \to \infty$, that means

$$p(k, t) \Rightarrow p_{st}(k) \quad \text{for} \quad t \to \infty. \tag{1.14}$$

1.3. Explicit choice of the transition probabilities; their relation to static random utility theory

Let us now make a simple and plausible explicit assumption about the form of the transition probability p_{ki}. It will turn out, that this form of the p_{ki} leads to a far-reaching analytic treatment of the master equation (1.9). The proposal reads

$$p_{ji} = \nu a_j b_i^{-1} \quad \text{with} \quad a_j > 0, b_i > 0 \quad \text{for} \quad j \neq i, \; p_{jj} \equiv 0. \tag{1.15}$$

The mobility factor ν (of dimension $[t^{-1}]$) can be absorbed into the time variable in (1.9) by introducing the dimensionless time

$$\tau = \nu t. \tag{1.16}$$

Therefore we shall henceforth leave away the factor ν and go over to the time variable τ.

Specializing to

$$a_i = b_i = \exp(u_i), \tag{1.17}$$

one obtains

$$p_{ji} = \exp(u_j - u_i), \qquad j \neq i. \tag{1.18}$$

In the present section we treat a_i and b_i as constants with time. In the following regression analysis however, the $u_i(t)$ can also be time dependent, and the proposal will be further generalized to

$$p_{ji} = v_{ji}\exp(u_j - u_i), \quad \text{and} \quad w_{ji} = n_i p_{ji} = n_i v_{ji}\exp(u_j - u_i), \tag{1.19}$$

where v_{ji} is a (symmetrical or even non-symmetrical) mobility matrix.

Coming back to the form (1.15), one can easily find the stationary solution of (1.9):

$$p_{st}(k) = c a_k b_k, \quad \text{with} \quad c = \left[\sum_{l=1}^{L} a_l b_l\right]^{-1}, \tag{1.20a}$$

or, inserting (1.17),

$$p_{st}(k) = c\exp(2u_k), \quad \text{with} \quad c = \left[\sum_{l=1}^{L} \exp(2u_l)\right]^{-1}. \tag{1.20b}$$

This corresponds to stationary mean regional population numbers

$$n_k^{st} = Nc\exp(2u_k). \tag{1.20c}$$

The proof follows by inserting (1.15) and (1.20a) into the right-hand side of (1.9):

$$\sum_{j=1}^{L} a_k b_j^{-1} c a_j b_j - \sum_{j=1}^{L} a_j b_k^{-1} c a_k b_k = c a_k \left\{\sum_{j=1}^{L} a_j - \sum_{j=1}^{L} a_j\right\} \equiv 0. \tag{1.21}$$

We shall now give an interpretation to the exponential factors u_i appearing in (1.18) and (1.20b) by comparing our *stationary* solution (1.20b) with the results of static random utility models.

These models come to the following result. If v_i is the 'utility' attributed to the alternative i (out of L alternatives $i = 1, 2, \ldots, L$), then the static (equilibrium) probability $P(k)$ of choosing the alternative k is given by

$$P(k) = \exp(v_k) \Big/ \sum_{l=1}^{L} \exp(v_l). \tag{1.22}$$

Since $P(k)$ corresponds to our *stationary* distribution $p_{st}(k)$, we have to identify

$$u_k = \tfrac{1}{2} v_k, \tag{1.23}$$

and hence to interpret the exponentials u_k in (1.20b) and (1.18) as 'utilities' (apart from the factor $\frac{1}{2}$). This interpretation also makes sense in the (not static but dynamic) concept of the transition probabilities (1.18): evidently, the individual transition probability p_{ji} from i to j is large (small), if the utility of the destination (origin) area exceeds the utility of the origin (destination) area.

The equilibrium state $p_{st}(k)$ of the dynamic system does not mean, that 'all utilities must be equal'. Instead, the dynamic equilibrium is a 'flux equilibrium'. In our case this flux equilibrium even fulfills the easily checked condition of *detailed balance*

$$p_{ji} p_{st}(i) = p_{ij} p_{st}(j), \tag{1.24}$$

which means, that in the stationary equilibrium state the flux of individuals from i to j is equal to the inverse flux from j to i, separately for each pair (i, j) of regions.

1.4. Explicit solution of the time dependent dynamic meanvalue equation

The content of our model is not exhausted by the above considerations, but goes far beyond the conclusions of the static random utility choice logit model (see also the end of section 4). The latter only determines a static distribution of the form (1.22), which corresponds to our stationary solution (1.20b), whereas we can find the evolution with time of probability distributions $p(k,t)$ with arbitrary non-equilibrium initial conditions.

In (1.13) we have seen, that only the conditional probability $p(k, \Delta t \mid i, 0)$ for an *infinitesimal* time interval Δt is directly proportional to $p_{kj} = a_k b_j^{-1}$ (for $k \neq j$). Now we have to determine the form of $p(k, t \mid i, 0)$ for *finite* time t. This amounts to summing up the effects of all transitions between regions which occur in the infinite sequence of the infinitesimal time intervals Δt. The Markov assumption and the Chapman–Kolmogorov equation are implied in this procedure.

Having found the conditional probability $p(k, t \mid i, 0)$, one can explicitly write down the evolution with time of the mean regional occupation numbers $\{n_1(t), n_2(t), \ldots, n_L(t)\}$ starting from arbitrary initial conditions $\{n_1(0), n_2(0), \ldots, n_L(0)\}$.

In solving the equation of motion (1.9) with transition probabilities (1.15) we proceed as follows. In section 1.4.1 we write the equation of motion in vector form and introduce equivalence transformations which prove useful in the further analysis. In section 1.4.2 the eigenvalues and eigenvectors are determined and in section 1.4.3 the conditional probability is expanded in terms of eigensolutions.

1.4.1. Equivalence transformations and eigensolutions of the meanvalue equations

It is convenient now to write eq. (1.9) (with dimensionless time variable τ) in vector notation:

$$\frac{\mathrm{d}\boldsymbol{P}(\tau)}{\mathrm{d}\tau}=L\boldsymbol{P}(\tau). \tag{1.25}$$

Here, $\boldsymbol{P}(\tau)$ is the L-dimensional vector with components $p(k,\tau)$, $k=1,2,\ldots,L$, and L is the matrix

$$L=((L_{ij})), \quad \text{with}$$

$$L_{ij}=p_{ij}-\delta_{ij}\sum_{k=1}^{L}p_{kj}=a_i b_j^{-1}-\delta_{ij}a\cdot b_j^{-1}, \quad \text{where} \tag{1.26}$$

$$a=\sum_{k=1}^{L}a_k.$$

Because of $L_{ij}\neq L_{ji}$ the matrix L is non-hermitean, that means $L^{+}\neq L$.

The eq. (1.25) can now be transformed into equivalent equations,

$$\frac{\mathrm{d}\boldsymbol{P}'(\tau)}{\mathrm{d}\tau}=L'(\tau)\boldsymbol{P}'(\tau), \tag{1.27}$$

by applying equivalence transformations with non-singular time independent matrices A:

$$\begin{aligned} L'&=A^{-1}LA, \qquad L=AL'A^{-1},\\ \boldsymbol{P}'&=A^{-1}\boldsymbol{P}, \qquad \boldsymbol{P}=A\boldsymbol{P}'. \end{aligned} \tag{1.28}$$

Two equivalence transformations of the kind (1.28) are of particular interest:

$$A\Rightarrow S=((\delta_{ij}p_{\mathrm{st},i}^{1/2}))=((\delta_{ij}(ca_i b_i)^{1/2})), \tag{1.29a}$$

$$\left.\begin{aligned} \tilde{L} &= ((\tilde{L}_{ij})) = S^{-1}LS = ((p_{\mathrm{st},i}^{-1/2} L_{ij} p_{\mathrm{st},j}^{1/2})) \\ &= ((a_i^{1/2} a_j^{1/2} b_i^{-1/2} b_j^{-1/2} - \delta_{ij} a b_i^{-1/2} b_j^{-1/2})), \\ \boldsymbol{Q} &= S^{-1}\boldsymbol{P}, \qquad \boldsymbol{P} = S\boldsymbol{Q}, \end{aligned}\right\} \tag{1.29b}$$

$$\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}\tau} = \tilde{L}\boldsymbol{Q}, \tag{1.29c}$$

and

$$A \Rightarrow B = ((\delta_{ij} b_i)), \tag{1.30a}$$

$$\left.\begin{aligned} \hat{L} &= ((\hat{L}_{ij})) = B^{-1}LB = ((b_i^{-1} L_{ij} b_j)) \\ &= ((b_i^{-1} a_i - \delta_{ij} a b_j^{-1})), \\ \boldsymbol{R} &= B^{-1}\boldsymbol{P}, \qquad \boldsymbol{P} = B\boldsymbol{R}, \end{aligned}\right\} \tag{1.30b}$$

$$\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}\tau} = \hat{L}\boldsymbol{R}. \tag{1.30c}$$

On the one hand, the transformation (1.29) leads to a *hermitean* (*real and symmetric*) *matrix* $\tilde{L} = \tilde{L}^+$, and, on the other hand, the transformation (1.30) leads to a particularly easily tractable form of the equations of motion.

The eigensolutions of (1.25), (1.29c) and (1.30c) are defined by

$$\boldsymbol{P}^{(\lambda)}(\tau) = \boldsymbol{P}^{(\lambda)} \mathrm{e}^{-\lambda\tau}, \quad \text{with} \quad L\boldsymbol{P}^{(\lambda)} = (-\lambda)\boldsymbol{P}^{(\lambda)}, \tag{1.31}$$

$$\boldsymbol{Q}^{(\lambda)}(\tau) = \boldsymbol{Q}^{(\lambda)} \mathrm{e}^{-\lambda\tau}, \quad \text{with} \quad \tilde{L}\boldsymbol{Q}^{(\lambda)} = (-\lambda)\boldsymbol{Q}^{(\lambda)}, \tag{1.32}$$

$$\boldsymbol{R}^{(\lambda)}(\tau) = \boldsymbol{R}^{(\lambda)} \mathrm{e}^{-\lambda\tau}, \quad \text{with} \quad \hat{L}\boldsymbol{R}^{(\lambda)} = (-\lambda)\boldsymbol{R}^{(\lambda)}. \tag{1.33}$$

The eigenvalues λ remain invariant under equivalence transformations of the type (1.28). Furthermore, it is known that for hermitean matrices like $\tilde{L}$ the system of eigenvectors $\boldsymbol{Q}^{(\lambda)}$ belonging to the eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{L-1}$ establishes a complete system of orthonormal vectors. Composing the L eigenvectors $\boldsymbol{Q}^{(\lambda_0)}, \boldsymbol{Q}^{(\lambda_1)}, \ldots, \boldsymbol{Q}^{(\lambda_{L-1})}$ with elements $q_i^{(\lambda)}$, $i = 1, 2, \ldots, L$ into an $L \times L$-matrix

$$Q = ((\boldsymbol{Q}^{(\lambda_0)}, \boldsymbol{Q}^{(\lambda_1)}, \ldots, \boldsymbol{Q}^{(\lambda_{L-1})})), \tag{1.34}$$

the orthonormality of the $Q^{(\lambda)}$ can be expressed by

$$Q^{+}\cdot Q=1, \quad \text{or,} \quad \text{in components,} \quad \sum_{i=1}^{L} q_i^{(\lambda)} q_i^{(\sigma)}=\delta_{\lambda\sigma}, \tag{1.35}$$

and the relation, which follows from (1.35),

$$Q\cdot Q^{+}=1, \quad \text{or,} \quad \text{in components,} \quad \sum_{\lambda=\lambda_0}^{\lambda_{L-1}} q_i^{(\lambda)} q_j^{(\lambda)}=\delta_{ij}, \tag{1.36}$$

expresses the completeness of the eigenvectors $Q^{(\lambda)}$.

From the properties of the eigenvectors $Q^{(\lambda)}$ of the hermitean matrix $\tilde{L}$ we can now conclude, that also the eigenvectors $P^{(\lambda)}$ of L and the eigenvectors $R^{(\lambda)}$ of $\hat{L}$ must establish *complete* systems of basis vectors, respectively, since the equivalence transformations S and B are non-singular. The systems of eigenvectors can easily be transformed into each other by making use of (1.29b) and (1.30b).

1.4.2. Determination of eigenvalues and eigenvectors

The explicit determination of eigenvectors and eigenvalues is implemented most easily by starting from the eigenvector eq. (1.33). Denoting the components of the eigenvector $R^{(\lambda)}$ by $r_i^{(\lambda)}$, this equation reads in components

$$(-\lambda)r_i^{(\lambda)}=b_i^{-1}a_i\sum_{j=1}^{L} r_j^{(\lambda)}-ab_i^{-1}r_i^{(\lambda)}, \quad \text{or} \tag{1.37a}$$

$$r_i^{(\lambda)}=\frac{a_i b_i^{-1} r^{(\lambda)}}{(ab_i^{-1}-\lambda)}, \quad \text{with} \quad r^{(\lambda)}=\sum_{j=1}^{L} r_j^{(\lambda)}. \tag{1.37b}$$

Summing (1.37b) over i and cancelling $r^{(\lambda)}$ one obtains

$$1=\sum_{i=1}^{L}\frac{a_i b_i^{-1}}{(ab_i^{-1}-\lambda)}. \tag{1.38}$$

This equation must be fulfilled by the eigenvalues λ.

Before proceeding further, we show that (in the non-degenerate case of differing b_i, which is only treated here) eq. (1.38) possesses exactly L solutions $\lambda_0, \lambda_1, \ldots, \lambda_{L-1}$. This is easily proved by graphic representation of the left-hand side and right-hand side of (1.38) as functions of λ in fig. 1.
We assume that the ab_i^{-1} are ordered according to

$$0<ab_1^{-1}<ab_2^{-1}<\cdots<ab_L^{-1}. \tag{1.39}$$

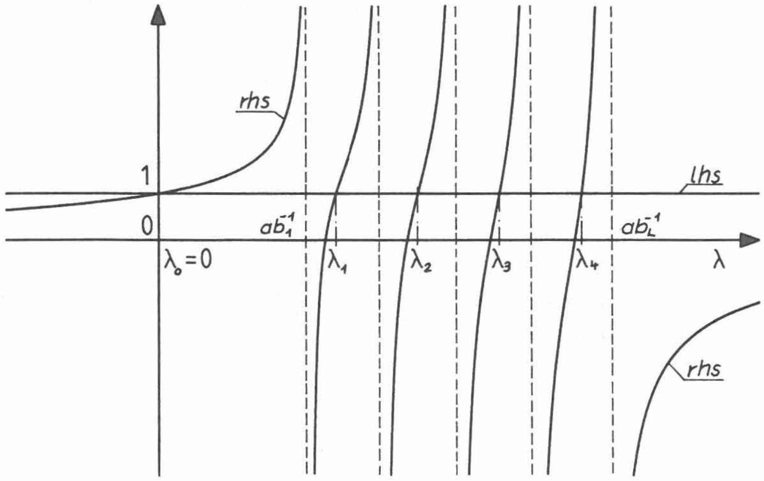


Fig. 1. Graphic representation of eq. (1.38).

The right-hand side of (1.38) is >0 for $\lambda<ab_1^{-1}$ and <0 for $\lambda>ab_L^{-1}$. It is equal to 1 for $\lambda=0$. Therefore, $\lambda_0=0$ is one of the eigenvalue solutions of (1.38), belonging to the stationary state, as we shall see. Furthermore, the right-hand side is singular ($\pm\infty$) at $\lambda=ab_j$, $j=1,2,\ldots,L$ because of the terms $\sim(ab_j^{-1}-\lambda)^{-1}$. The function passes from $-\infty$ to $+\infty$ in each interval $[ab_j^{-1}, ab_{j+1}^{-1}]$ and we find exactly one eigenvalue in each interval:

$$ab_j^{-1}<\lambda_j<ab_{j+1}^{-1}, \qquad j=1,2,\ldots,L-1. \tag{1.40}$$

Hence, with the exception of $\lambda_0=0$, all eigenvalues λ_j are positive. The exact values of the λ_j can only be found by numerical evaluation of (1.38). These values are however not necessary for establishing the general form of the time-dependent solution.

The form of the eigenvectors $\boldsymbol{P}^{(\lambda)}$, $\boldsymbol{Q}^{(\lambda)}$ and $\boldsymbol{R}^{(\lambda)}$ with components $p_i^{(\lambda)}$, $q_i^{(\lambda)}$ and $r_i^{(\lambda)}$, respectively, now follows from (1.37b) and the equivalence transformations (1.29b), (1.30b).

$$p_i^{(\lambda)}=b_i r_i^{(\lambda)}=\frac{g(\lambda)a_i}{(ab_i^{-1}-\lambda)}, \tag{1.41}$$

$$q_i^{(\lambda)}=p_{\mathrm{st},i}^{-1/2}p_i^{(\lambda)}=(ca_ib_i)^{-1/2}b_ir_i^{(\lambda)}=\frac{g(\lambda)c^{-1/2}a_i^{1/2}b_i^{-1/2}}{(ab_i^{-1}-\lambda)}, \tag{1.42}$$

$$r_i^{(\lambda)}=\frac{g(\lambda)a_ib_i^{-1}}{(ab_i^{-1}-\lambda)}. \tag{1.43}$$

The factor $g^{(\lambda)}$ follows from the correct normalization of the $\boldsymbol{Q}^{(\lambda)}$:

$$1=\|\boldsymbol{Q}^{(\lambda)}\|^2=\sum_{i=1}^{L} q_i^{(\lambda)}q_i^{(\lambda)}=\frac{g^2(\lambda)}{c}\sum_{i=1}^{L}\frac{a_i b_i^{-1}}{(ab_i^{-1}-\lambda)^2}, \quad \text{or}$$
$$g(\lambda)=c^{1/2}\left(\sum_{i=1}^{L}\frac{a_i b_i^{-1}}{(ab_i^{-1}-\lambda)^2}\right)^{-1/2}. \tag{1.44}$$

In particular it can be checked in (1.41), that the eigenmode $\boldsymbol{P}^{(\lambda_0)}$ belonging to $\lambda_0=0$ has the components

$$p_k^{(\lambda_0=0)}=\frac{g(0)}{a}a_k b_k\equiv c a_k b_k, \tag{1.45}$$

and is identical with the stationary solution $p_{st}(k)$ [see (1.20a)].

1.4.3. Representation of time dependent solutions of the meanvalue equations

Any time dependent probability distribution, which is a solution of (1.9) with transition probabilities (1.15) can now be expanded as a linear combination of the eigensolutions (1.31),

$$\boldsymbol{P}(\tau)=\sum_{\lambda=\lambda_0}^{\lambda_{L-1}}\boldsymbol{P}^{(\lambda)}\mathrm{e}^{-\lambda\tau}c^{(\lambda)}, \tag{1.46a}$$

or, in components,

$$p(j;\tau)=\sum_{\lambda=\lambda_0}^{\lambda_{L-1}}p_j^{(\lambda)}\mathrm{e}^{-\lambda\tau}c^{(\lambda)}, \tag{1.46b}$$

where (1.41) has to be inserted.

This holds in particular for the conditional probability

$$p(j,\tau\,|\,i,0)=\sum_{\lambda=\lambda_0}^{\lambda_{L-1}}p_j^{(\lambda)}\mathrm{e}^{-\lambda\tau}c_i^{(\lambda)}, \tag{1.47}$$

where now the expansion coefficients $c_i^{(\lambda)}$ depend on the initial state i. In order to fulfill the initial condition (1.10a), we must choose

$$c_i^{(\lambda)}=q_i^{(\lambda)}p_{st,i}^{-1/2}=q_i^{(\lambda)}(c a_i b_i)^{-1/2}. \tag{1.48}$$

Indeed, with (1.47) and (1.48) there follows for $\tau=0$,

$$p(j,0\,|\,i,0)=\sum_{\lambda=\lambda_0}^{\lambda_{L-1}} p_j^{(\lambda)}q_i^{(\lambda)}p_{\mathrm{st},i}^{-1/2}$$

$$=\sum_{\lambda=\lambda_0}^{\lambda_{L-1}} p_{\mathrm{st},j}^{1/2}q_j^{(\lambda)}q_i^{(\lambda)}p_{\mathrm{st},i}^{-1/2}=p_{\mathrm{st},j}^{1/2}\,\delta_{ji}p_{\mathrm{st},i}^{-1/2}=\delta_{ji}. \tag{1.49}$$

Here we have made use of the completeness relation (1.36). On the other hand, also (1.10b) is fulfilled for all times τ, because of

$$\sum_{j=1}^{L} p_j^{(\lambda)}=\sum_{j=1}^{L} q_j^{(\lambda)}p_{\mathrm{st},j}^{1/2}=\sum_{j=1}^{L} q_j^{(\lambda)}q_j^{(\lambda_0)}=\delta_{\lambda\lambda_0} \tag{1.50}$$

[see the orthogonality relation (1.35)], and because of

$$c_i^{(\lambda_0)}=q_i^{(\lambda_0)}(ca_ib_i)^{-1/2}=1. \tag{1.51}$$

Furthermore, we see that for $\tau\to\infty$ the conditional probability $p(j,\tau\,|\,i,0)$ approaches the stationary solution

$$\lim_{\tau\to\infty} p(j,\tau\,|\,i,0)=p_j^{(\lambda_0)}=p_{\mathrm{st},j}=ca_jb_j, \tag{1.52}$$

independently of the initial state i! This follows, because $\lambda_j>0$ for $j=1,2,\ldots,L-1$; therefore all exponential factors $\mathrm{e}^{-\lambda_j\tau}$ die out with the exception of $\mathrm{e}^{-\lambda_0\tau}\equiv 1$. According to (1.11) the conditional probability (1.47) can now be used to construct time dependent solutions of the meanvalue eq. (1.9) with *arbitrary* initial conditions.

As the Chapman–Kolmogorov equation (1.12) is constitutive for the whole probability theory of time dependent Markov-processes, we finally prove, that this relation is satisfied by the conditional probability (1.47).

Writing $p(j,\tau_2\,|\,i,\tau_1)$ and $p(k,\tau_3\,|\,j,\tau_2)$ in the form

$$\begin{aligned} p(j,\tau_2\,|\,i,\tau_1)&=\sum_{\lambda=\lambda_0}^{\lambda_{L-1}} \mathrm{e}^{-\lambda(\tau_2-\tau_1)}p_{\mathrm{st},j}^{1/2}q_j^{(\lambda)}q_i^{(\lambda)}p_{\mathrm{st},i}^{-1/2},\\ p(k,\tau_3\,|\,j,\tau_2)&=\sum_{\lambda'=\lambda_0}^{\lambda_{L-1}} e^{-\lambda'(\tau_3-\tau_2)}p_{\mathrm{st},k}^{1/2}q_k^{(\lambda')}q_j^{(\lambda')}p_{\mathrm{st},j}^{-1/2}, \end{aligned} \tag{1.53}$$

there follows, making use of the orthonormality relation (1.35),

$$\sum_{j=1}^{L} p(k,\tau_3 \mid j,\tau_2) p(j,\tau_2 \mid i,\tau_1)$$

$$= \sum_{j=1}^{L} \sum_{\lambda=\lambda_0}^{\lambda_{L-1}} \sum_{\lambda'=\lambda_0}^{\lambda_{L-1}} e^{-\lambda'(\tau_3-\tau_2)-\lambda(\tau_2-\tau_1)} p_{st,k}^{1/2} q_k^{(\lambda')} q_j^{(\lambda')} q_j^{(\lambda)} q_i^{(\lambda)} p_{st,i}^{-1/2}$$

$$= \sum_{\lambda=\lambda_0}^{\lambda_{L-1}} \sum_{\lambda'=\lambda_0}^{\lambda_{L-1}} e^{-\lambda'(\tau_3-\tau_2)-\lambda(\tau_2-\tau_1)} p_{st,k}^{1/2} q_k^{(\lambda')} \delta_{\lambda'\lambda} q_i^{(\lambda)} p_{st,i}^{-1/2}$$

$$= \sum_{\lambda=\lambda_0}^{\lambda_{L-1}} e^{-\lambda(\tau_3-\tau_1)} p_{st,k}^{1/2} q_k^{(\lambda)} q_i^{(\lambda)} p_{st,i}^{-1/2} = p(k,\tau_3 \mid i,\tau_1). \tag{1.54}$$

This proves the Chapman–Kolmogorov relation.

2. Determination of mobilities and utilities from empirical data

After having worked out the theoretical framework of a dynamic migration theory we have now to compare the theory with empirical data. We attribute a utility u_i to each region i and also assume mobilities v_{ji} between region i and j, which will depend on the spatial distance and other factors. This amounts to the specialization (1.17) of the factors a_i, b_i and to the proposal (1.19) for the migration matrix w_{ji}. In contrast to section 1, where we assumed constant utilities u_i and factors a_i, b_i, we do now allow the utilities u_i (and hence the transition probabilities) possibly to become functions of time τ. The framework of equations of motion however remains the same under this generalization.

On the other hand, in a migration system with L areas and one population, the following *empirical* quantities listed in table 1 can be observed year by year.

Table 1
Observed quantities per year – describing the migration process.

Area	Population size	Number of transitions per year from i to j							
		(1)	(2)	(3)	(·)	(·)	(j)	(·)	(L)
1	n_1^e	*	w_{21}^e	w_{31}^e	…	…	w_{j1}^e	…	w_{L1}^e
2	n_2^e	w_{12}^e	*	w_{32}^e	…	…	w_{j2}^e	…	w_{L2}^e
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	n_i^e	w_{1i}^e	w_{2i}^e	w_{3i}^e	…	…	w_{ji}^e	…	w_{Li}^e
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
L	n_L^e	w_{1L}^e	w_{2L}^e	w_{3L}^e	…	…	w_{jL}^e	…	*

The *theoretical* migration matrix introduced in (1.19),

$$w_{ji}(v,u)=n_i v_{ji}\exp(u_j-u_i), \tag{2.1}$$

is a function of the mobility v_{ji} and the utilities u_i and u_j. Now, it has to be matched to the *empirical* migration matrix $w^e_{ji}(t)$ by an optimal estimation of the mobilities v_{ji} and utilities u_i, u_j.

The L^2-L observations of the $w^e_{ji}(t)$, that is the numbers of migrants per year from any area i to any area $j(\neq i)$, enable us to estimate the unknown values of the utilities. The estimation is implemented in two steps of increasing generality. In section 2.1 we assume a symmetric mobility matrix $v_{ji}=v_{ij}$. In section 2.2 we generalize the estimation procedure for an asymmetric transportation cost or 'distance'-matrix $d_{ij}\neq d_{ji}$, which leads to an asymmetric mobility matrix $v_{ji}\neq v_{ij}$.

2.1. *Estimation of utilities for a symmetric mobility matrix*

In this section we assume a symmetric mobility matrix

$$v_{ij}=v_{ji}. \tag{2.2}$$

The meaning of (2.2) can be seen from considering two regions i, j with equal utilities $u_i=u_j$ and equal population numbers $n_i=n_j$. It follows from (2.1), with (2.2) presumed, that in this case the number of transitions from i to j is equal to that from j to i. This implies a symmetry of transportation costs for $i\to j$ and $j\to i$.

Let us now first consider the product $w_{ij}\cdot w_{ji}$. Using (2.1) with (2.2) and equating theoretical with empirical expressions we obtain the products

$$w^e_{ij}(t)\cdot w^e_{ji}(t)=w_{ij}(v,u)\cdot w_{ji}(v,u)=n_i n_j v^2_{ij} \tag{2.3}$$

yielding the mobility parameters

$$v_{ij}(t)=\sqrt{\frac{w^e_{ij}(t)\cdot w^e_{ji}(t)}{n^e_i(t)\cdot n^e_j(t)}}>0. \tag{2.4}$$

Since interregional mobility parameters turn out to have only small fluctuations around a mean value (e.g., for the Canadian system), we later insert into the simulation procedure their temporal mean value

$$v_{ij}=\frac{1}{T}\sum_{t=1}^{T} v_{ij}(t). \tag{2.5}$$

Furthermore it is possible to define a global mobility parameter v characterizing the mean mobility of the population under consideration,

$$v(t)=\frac{1}{L(L-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{L} v_{ij}(t). \tag{2.6}$$

We emphasize, that the determination of the mobility parameters v_{ij} is *independent* of the estimation of the utilities u_i.

Second, we consider the quotient w_{ij}/w_{ji}. Using (2.1) with (2.2) and equating empirical with theoretical expressions, we obtain for the quotients

$$\frac{w_{ij}^{e}(t)}{w_{ji}^{e}(t)}=\frac{w_{ij}(v,u)}{w_{ji}(v,u)}=\frac{n_j^{e}}{n_i^{e}}\exp[2(u_i-u_j)], \tag{2.7}$$

or, equivalently,

$$C_{ij}(t)=u_i(t)-u_j(t), \qquad i,j=1,\dots,L, \quad i\neq j, \tag{2.8}$$

where we have introduced the antisymmetric matrix

$$C_{ij}(t)=\tfrac{1}{2}\ln\frac{w_{ij}^{e}(t)n_i^{e}(t)}{w_{ji}^{e}(t)n_j^{e}(t)}=-C_{ji}(t). \tag{2.9}$$

Eqs. (2.8) are linear in the utilities to be determined. Since we have (L^2-L) equations but only L unknown utilities, the utilities in general cannot be chosen to satisfy (2.8) exactly. We can, however, choose an optimal set of utilities by minimizing the least square deviations

$$F_t[\boldsymbol{u}]=\sum_{\substack{i,j=1\\ i\neq j}}^{L}[C_{ji}(t)-(u_j(t)-u_i(t))]^2. \tag{2.10}$$

This yields the optimal utilities

$$u_i(t)=\frac{1}{L}\sum_{\substack{j=1\\ j\neq i}}^{L} C_{ij}(t)+\frac{1}{L}\sum_{j=1}^{L} u_j(t), \tag{2.11}$$

where the sum $\sum_{j=1}^{L} u_j(t)$ remains undetermined by the optimization procedure. We can however put

$$\sum_{j=1}^{L} u_j(t)=0 \tag{2.12}$$

in (2.11) without changing the values $w_{ij}(v,u)$ of migration matrix (2.1). This concludes the determination of the $u_i(t)$.

2.2. Estimation of mobilities and utilities for an asymmetric transportation cost matrix

The mobility matrix v_{ij} introduced above is related to the concept of 'distance' between the areas or regions i and j. 'Distance' is one of the most important specific variables in spatial analysis, and many possibilities for its analytical specification are offered. However, the concept of 'distance' must be generalized to comprise geographical and economic aspects.

Geographically the length of routes between places i, j can be used as an appropriate measure d_{ij} of distance. If the same route is used in both directions, we find $d_{ij}=d_{ji}$.

Economic distances, also denoted by d_{ij}, are measured in terms of costs that can be evaluated in time or money. Traffic congestion can be important here. In general, the transport-cost–distance matrix will be asymmetric, that is $d_{ij} \neq d_{ji}$.

For constructive specification, we assume the following reasonable relation between 'distance' d_{ij} and mobility $v_{ij}(t)$:

$$v_{ij}(t) = v(t) \cdot \exp[-\beta(t) d_{ij}], \tag{2.13}$$

where in general $d_{ij} \neq d_{ji}$, hence $v_{ij} \neq v_{ji}$. Inserting (2.13) into (2.1), we obtain

$$w_{ij}(v,u) = v n_j \cdot \exp[-\beta d_{ij}] \exp[u_i - u_j]. \tag{2.14}$$

The global mobility parameter $v(t)$ and the trendparameter $\beta(t)$ which is a measure of the deterrence of the costs of travel will now be estimated in analogy to our procedure in section 2.1.

First, we consider the product $w_{ij} \cdot w_{ji}$. Using (2.14) we now obtain, generalizing (2.3),

$$w_{ij}^{e}(t) \cdot w_{ji}^{e}(t) = w_{ij}(v,u) \cdot w_{ji}(v,u) = v^2 n_i^e n_j^e \exp[-\beta(d_{ij}+d_{ji})],$$

and therefore, after taking the logarithm,

$$S_{ij}(t) = \mu - \beta \frac{(d_{ij}+d_{ji})}{2}, \tag{2.15}$$

where we have introduced

$$S_{ij}(t)=S_{ji}(t)=\tfrac{1}{2}\ln\frac{w_{ij}^{e}(t)w_{ji}^{e}(t)}{n_i^{e}(t)n_j^{e}(t)},\qquad i\neq j,\quad\text{and}\tag{2.16}$$

$$\mu=\ln \nu.\tag{2.17}$$

We now estimate the parameters μ, β by minimizing the functional

$$G_t[\mu,\beta]=\sum_{\substack{i,j=1\\ i\neq j}}^{L}\left[S_{ij}(t)-\left(\mu-\beta\frac{(d_{ij}+d_{ji})}{2}\right)\right]^2,\tag{2.18}$$

according to

$$\partial G_t/\partial\mu=0,\qquad \partial G_t/\partial\beta=0.\tag{2.19}$$

Eqs. (2.19) yield the results

$$\beta(t)=\frac{\overline{S(t)}\cdot\overline{d(t)}-\overline{Sd(t)}}{(\overline{d^2(t)}-\bar{d}^2(t))},\quad\text{and}\tag{2.20a}$$

$$\mu(t)=\bar{S}(t)+\bar{d}(t)\cdot\beta(t),\tag{2.20b}$$

with the abbreviations

$$\begin{aligned}
\bar{S}(t)&=\frac{1}{L(L-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{L}S_{ij}(t),\\
\bar{d}(t)&=\frac{1}{L(L-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{L}\frac{(d_{ij}(t)+d_{ji}(t))}{2},\\
\overline{Sd(t)}&=\frac{1}{L(L-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{L}S_{ij}(t)\cdot\frac{(d_{ij}(t)+d_{ji}(t))}{2},\\
\overline{\mathrm{d}^2(t)}&=\frac{1}{L(L-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{L}\left(\frac{d_{ij}(t)+d_{ji}(t)}{2}\right)^2.
\end{aligned}\tag{2.21}$$

By (2.20) the deterrence parameter $\beta(t)$ and the global mobility parameter $\nu(t)$ are determined. The estimation procedure for these parameters is again independent of the estimation of the utilities.

Second, we consider the quotient w_{ij}/w_{ji}. Generalizing (2.7), we obtain

$$\frac{w_{ij}^{e}(t)}{w_{ji}^{e}(t)}=\frac{w_{ij}(v,\tilde{u})}{w_{ji}(v,\tilde{u})}=\frac{n_j^e}{n_i^e}\exp[-\beta(d_{ij}-d_{ji})]\exp[2(\tilde{u}_i-\tilde{u}_j)], \tag{2.22}$$

or, after taking the logarithm,

$$\tilde{C}_{ij}(t)=\tilde{u}_i-\tilde{u}_j, \qquad i,j=1,2,\ldots,L, \quad i\neq j, \quad \text{with} \tag{2.23}$$

$$\tilde{C}_{ij}(t)=C_{ij}(t)+\frac{\beta}{2}(d_{ij}-d_{ji})=-\tilde{C}_{ji}. \tag{2.24}$$

According to section 2.1 the optimal set of utility functions can be estimated. This results in

$$\tilde{u}_i(t)=\frac{1}{L}\sum_{\substack{j=1\\ j\neq i}}^{L}\tilde{C}_{ij}(t)$$

or, with (2.11),

$$\tilde{u}_i(t)=u_i(t)+\frac{\beta(t)}{2L}\sum_{\substack{j=1\\ j\neq i}}^{L}(d_{ij}-d_{ji}). \tag{2.25}$$

We see, that the asymmetric part of the transport-cost–distance matrix d_{ji} yields a modification of the original utilities for symmetric mobility.

3. The dependence of utilities and mobilities on motivation factors

The estimation procedure, described in section 2 yields with some accuracy the time series for the parameters $\tau(t)$, where $\tau(t)\in\{v(t),\beta(t),u_i(t),\ldots\}$. These trend parameters in their turn determine the dynamics of the migration system. Therefore, it is promising to correlate the trend parameters $\tau(t)$ to a set of exogenous key-factors $\mu_s(t)$, $s=1,2,\ldots,S$. These sets of key-factors are properly standardized and detrended, in the time interval $0<t<T$ under consideration. A comparative study of different nations could distinguish between *international* and *national* key-factors and events. A tentative relation between the trend parameters and the key-factors is assumed as

$$\tau^{\text{th}}(t)=\sum_{s=1}^{S}a_s\mu_s(t-t_s), \qquad t=1,2,\ldots,T, \tag{3.1}$$

where we have introduced an eventual time-lag t_s. The weight factors a_s of the different key-factors can be obtained using again least-square estimation,

$$F_t[a_1,\dots,a_s]=\sum_{t=1}^{T}\left[\tau^e(t)-\sum_{s=1}^{S}a_s\mu_s(t-t_s)\right]^2=\text{Min!} \tag{3.2}$$

This procedure results in a system of linear equations which have to be solved,

$$\sum_{s=1}^{S}a_sC_{rs}=D_r, \qquad r=1,2,\dots,S, \quad \text{with} \tag{3.3}$$

$$C_{sr}=C_{rs}=\frac{1}{T}\sum_{t=1}^{T}\mu_s(t-t_s)\mu_r(t-t_r), \quad \text{and} \tag{3.4}$$

$$D_T=\frac{1}{T}\sum_{t=1}^{T}\tau(t)\mu_r(t-t_r). \tag{3.5}$$

A measure for the agreement between $\tau^{th}(t)$ and $\tau^e(t)$ is the corresponding correlation coefficient. Key-factors for the present analysis may be economic, socio-economic or population-variables – or powers of them – e.g., income per capita, number of open positions in industry, housing market, differences in the tax situation, population densities, interaction between different populations.

4. The Canadian interregional migration system

In this section we apply our theory to the Canadian interregional migration system. We present only a few results, without going into detail. The results will be further evaluated and correlated to socio-economic data. Fig. 2 represents the map of Canada with eleven regions. The regions are the following.

(1) New Foundland	(7) Manitoba
(2) Prince Edward Island	(8) Saskatchewan
(3) Nova Scotia	(9) Alberta
(4) New Brunswick	(10) British Columbia
(5) Quebec	(11) Yukon and Northwest Territories
(6) Ontario	

In fig. 3 we plotted using (2.11) the optimal utility functions $u_i(t)$ for the different Canadian regions. The robustness of all parameters (utilities, mobilities) with respect to data uncertainties could be established. Preferred regions are (5, 6, 9, 10), more or less neutral regions (1, 3, 4, 7, 8) and not

Fig. 2. Map of Canada.

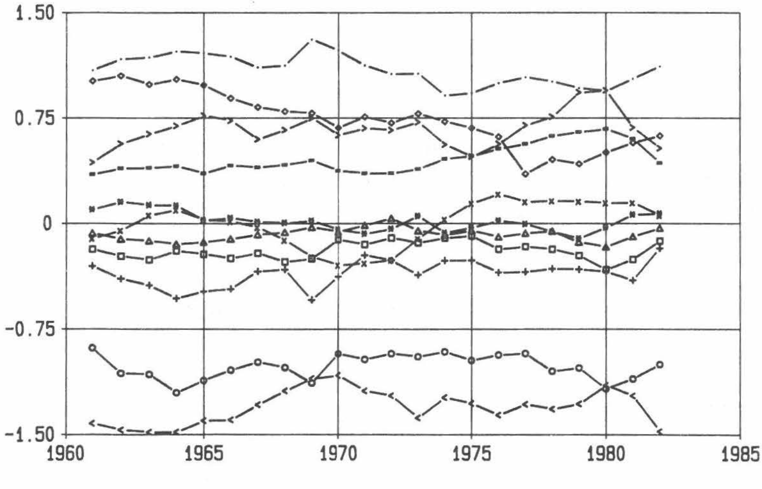


Fig. 3. Utilities of Canada $u_i(t)$.

New Foundland	+	Manitoba	*
Prince Edward Island	○	Saskatchewan	×
Nova Scotia	△	Alberta	=
New Brunswick	□	British Columbia	>
Quebec	◇	Yukon and	
Ontario	·	Northwest Territories	<

preferred regions (2, 11). In section 3, it is shown how the time series of the utilities can be linked to variations of certain motivation factors $\mu_s(t)$.

In fig. 4 the spatial variance

$$\sigma_L^2(t)=\frac{1}{L}\sum_{i=1}^{L} u_i^2(t)-\left(\frac{1}{L}\sum_{i=1}^{L} u_i(t)\right)^2 \tag{4.1}$$

of the utilities as a function of time is plotted.

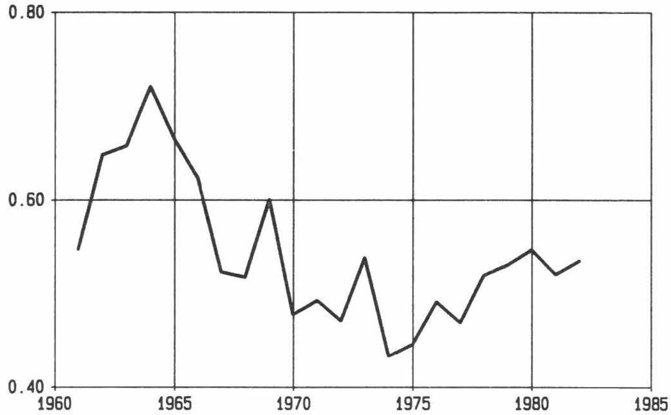


Fig. 4. Spatial variance of the utilities.

In fig. 5 the global mobility factor $\nu(t)$ after (2.6) is depicted.

For each year the estimated set of utilities $u_i(t)$ can be used to determine the *virtual stationary* mean occupation numbers

$$n_k^{st}(t)=Nc(t)\cdot\exp(2u_k(t)), \qquad c(t)=\left(\sum_{i=1}^{L}\exp(2u_i(t))\right)^{-1}. \tag{4.2}$$

These theoretical quantities describe the equilibrium occupation numbers, into which the system *would evolve* in the future, if the utilities $u_k(t)$ would remain constant from time t onward until $t\rightarrow\infty$. These virtual stationary occupation numbers are in general not realized at time t, since the system is not in equilibrium with respect to the momentary utilities $u_k(t)$. Instead, the actually realized occupation numbers $\{n_1^e(t),\dots,n_L^e(t)\}$ will in general differ from the $\{n_1^{st}(t),\dots,n_L^{st}(t)\}$ defined by (4.2). The deviation of the $\boldsymbol{n}^e(t)$ from $\boldsymbol{n}^{st}(t)$ describes the 'distance from equilibrium' and can be seen as defining the 'migratory stress' in the population. A compact measure for the migratory

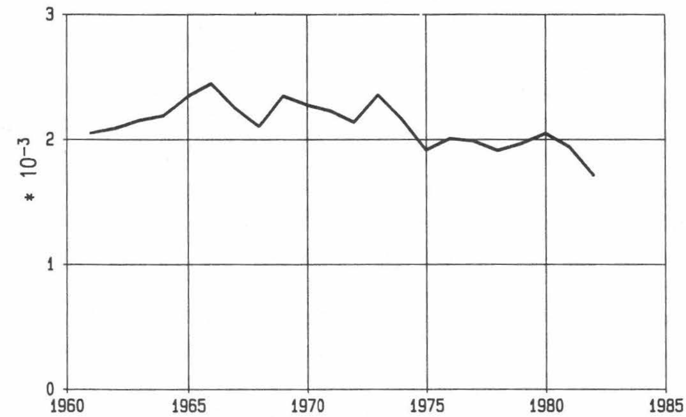


Fig. 5. Global mobility of Canada.

stress can be derived from the correlation coefficient $r_t(\boldsymbol{n}^{\mathrm{e}}, \boldsymbol{n}^{\mathrm{st}})$ between the actual empiric $\boldsymbol{n}^{\mathrm{e}}(t)$ and the virtual stationary $\boldsymbol{n}^{\mathrm{st}}(t)$. Obviously $r_t(\boldsymbol{n}^{\mathrm{e}}, \boldsymbol{n}^{\mathrm{st}}) \Rightarrow 1$ for $\boldsymbol{n}^{\mathrm{e}} \Rightarrow \boldsymbol{n}^{\mathrm{st}}$. Therefore $(1 - r_t(\boldsymbol{n}^{\mathrm{e}}, \boldsymbol{n}^{\mathrm{st}}))$ measures the migratory stress. In fig. 6 the correlation coefficient $r_t(\boldsymbol{n}^{\mathrm{e}}, \boldsymbol{n}^{\mathrm{st}})$ is shown for the Canadian system. Today Canada is more distant from its virtual migratory equilibrium than in the sixties.

The consideration of the 'migratory stress' finally can be used to make clear the difference between our model and a 'repeated static utility choice logit model'.

The latter would proceed as follows. From time to time it would take the occupation numbers $\{n_1(t), \ldots, n_L(t)\}$ and determine the utilities $\{u_1(t), \ldots, u_L(t)\}$

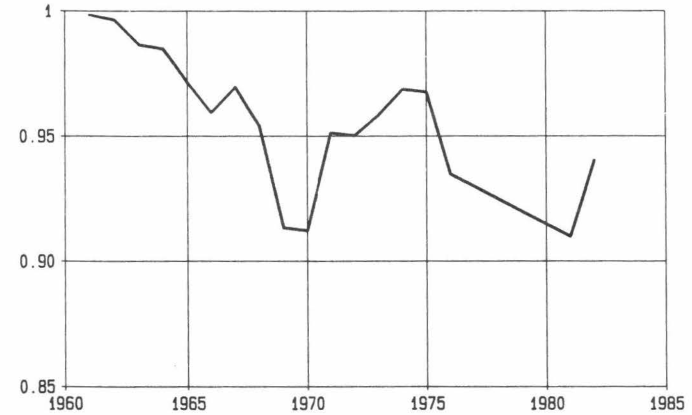


Fig. 6. Correlation coefficient $r(\boldsymbol{n}^{\mathrm{e}}, \boldsymbol{n}^{\mathrm{st}})$.

belonging to them by making use of formula (4.2). This, however, amounts to the (unjustified) identification of the real momentary occupation numbers $\{n_1(t), \ldots, n_L(t)\}$ with the virtual stationary occupation numbers $\{n_1^{st}(t), \ldots, n_L^{st}(t)\}$ which the system would reach after sufficient relaxation time, if the utilities would remain so. Such a model could only be justified, if the relaxation time to reach the equilibrium belonging to the utilities $\{u_1(t), \ldots, u_L(t)\}$ would be infinitely short. Only then the empiric $\boldsymbol{n}^e(t)$ would agree with the $\boldsymbol{n}^{st}(t)$. Then the correlation coefficient $r_t(\boldsymbol{n}^e, \boldsymbol{n}^{st})$ would always be equal to 1, and no migratory stress would exist. Fig. 6 shows that this is not the case for real systems.

Our procedure is different. We do not assume that the momentary $\boldsymbol{n}^e(t)$ are in (quasistatic) equilibrium with the momentary utilities $\{u_1(t), \ldots, u_L(t)\}$. Instead we say that the momentary utilities only determine the momentary transition probabilities, which in turn determine – via the equations of motion – the further evolution of $\{n_1(t), \ldots, n_L(t)\}$ with time. Vice versa, the utilities are found from the transition probabilities via the regression analysis of section 2.

References

Haag, G. and W. Weidlich, 1984, A stochastic theory of interregional migration, Geographical Analysis 16, 331–357.

Weidlich, W. and G. Haag, n.d., A dynamic phase transition model for spatial agglomeration processes, Journal of Regional Science, forthcoming.

Regional Science and Urban Economics 16 (1986) 81–95. North-Holland

MULTIPLIER–ACCELERATOR MODELS REVISITED

Tönu PUU

University of Umeå, S-901 87 Umeå, Sweden

Received April 1985, final version received August 1985

This article discusses multiplier–accelerator models in continuous time as formulated once by Phillips. The model is placed in a continuous two-dimensional spatial setting and augmented by an interregional trade multiplier. It is seen that this produces complex patterns of change like a distributed lag system. Moreover, the action of the accelerator is made non-linear as suggested by Goodwin. This changes the system to produce limit cycles. The limit cycles may bifurcate due to parametric change. In the spatial setting it is seen that the space coordinates can cause bifurcation of the temporal limit cycle pattern.

1. Introduction

The objective of the present paper is to extend the multiplier–accelerator model of business cycles in two directions. *First*, it is placed in a continuous two-dimensional spatial setting with interregional trade generated by a linear propensity to import. *Second*, the idea of a floor and a ceiling, as suggested by Hicks (1950) to limit the action of the linear accelerator, is incorporated into the model in terms of a non-linear investment function. The way this is done much resembles that due to Goodwin (1951), except that in our case no mixed difference-differential equation results. The analysis is devoted to the simple Samuelson (1939)–Hicks (1950) model in its elegant continuous shape cast by Phillips (1954).

The conclusion from introducing non-linearity is the same as Goodwin's. Instead of explosion or extinction of motion as alternatives we get persistent bounded limit cycles. However, it is possible to replace Goodwin's general reasoning by results from singular perturbation theory developed by 1970 in applied mathematics. Moreover, recent developments in bifurcation and catastrophe theory make it possible to analyse bifurcations of limit cycles for only slightly more complicated investment functions than those used by Goodwin.

The introduction of space is shown to destroy the perfect periodicity of the original model, and to replace its simple harmonic motion by irregular time profiles. In the latter aspect space acts like a distributed lag system. In the first it does something distributed lags cannot do, i.e., compound cycles of non-rational period relations.

The combination of non-linearity with a spatial setting gives two main results. A general theorem on the impossibility of spatially homogeneous steady state solutions is proved. Further the model is completely solved for the case of a dispersive wave proceeding radially from the origin. It is shown that the spatial coordinate (radius vector) itself acts as a parameter causing bifurcation from an area where motion vanishes to one where limit cycles persist.

2. The original model

The multiplier–accelerator model was formulated by Samuelson (1939), and was later elaborated by Hicks (1950). Unlike the contemporary growth models the business cycle models were formulated as difference equations in discrete time. The choice between a discrete or continuous formulation is always one of convenience. For numerical computation or empirical study the discrete form is necessary. For analytical purposes a continuous representation of time (and space) is always superior. Therefore, the starting point will be a multiplier–accelerator model formulated by Phillips as an adaptive process in continuous time. A full account of this model can be found in Allen (1956).

The essential elements of the model are saving (or consumption) and 'induced' investments. We disregard the so called autonomous expenditures, as they only introduced a non-zero equilibrium income from which the solution of the autonomous model yields the deviations. As the particular and homogeneous solutions are additive this makes no harm, at least in the linear case, but saves unnecessary symbolism.

Savings are assumed to be a given proportion, s, of income, Y. Induced investments are proportionate to the rate of change of income, $\dot{Y}$, the proportionality factor (the accelerator) being denoted v. Obviously, v is the fixed input coefficient of capital in a Leontief type of production function where labour is affluent and capital is always the limiting factor. Equating savings to investments, $sY=v\dot{Y}$, yields the Harrod model of balanced growth. Corresponding to the adjustment delay in the Samuelson discrete model the continuous model has adaptive adjustment, so that the rate of change of income, $\dot{Y}\propto(I-sY)$, the difference between current investment and savings, and so that the rate of change of investments, $\dot{I}\propto(v\dot{Y}-I)$, the difference between 'optimal' investments in view of income change and actual investments. The model becomes particularly simple if we, as do Samuleson and Hicks with the lagged delays, assume that the proportionality constants in these adjustment processes are equal and can be made unitary by a suitable choice of time unit. So, $\dot{Y}=(I-sY)$ and $\dot{I}=(v\dot{Y}-I)$, from which I and $\dot{I}$ can be eliminated by differentiating the first equation once more with respect to time and then using the original pair of equations for the elimination. Accordingly,

$$\ddot{Y}+(1+s-v)\dot{Y}+sY=0 \tag{1}$$

is obtained. Its solutions are like those of the original Samuelson–Hicks model formulated for discrete time, i.e., they are those of a simple damped or antidamped harmonic oscillator of one definite period.

3. Space and interregional trade

We now place the model in the setting of continuous two-dimensional space and introduce exports X and imports M along with the variables Y and I. It seems logical to assume a linear export–import multiplier along with the linear multiplier for local consumption. Denoting the propensity to import by m, imports are proportional to local income and exports to income 'abroad'. Accordingly, $X-M$ in each location is equal to m multiplied by income difference abroad and at home. Assuming only local action between spatially contiguous locations (a first approximation equivalent to the assumption that complicated lag structures are absent in time), we still have to formalize how spatial income differences should be measured.

As demonstrated by the author in Puu (1982) and in Beckmann and Puu (1985) by Gauss' integral theorem the proper measure of spatial income differences is the Laplacian of income $\nabla^2 Y=\partial^2 Y/\partial x^2+\partial^2 Y/\partial y^2$, where x and y are the euclidean coordinates of twospace. Thus $X-M=m\nabla^2 Y$, or in adaptive form $\dot{X}-\dot{M}\propto(m\nabla^2 Y-X+M)$. Again assuming that adjustment speed is the same as the two previous ones, we can replace the proportionality sign by an equality.

As export surplus enters the system in the same way as do induced investments, the relations become $\dot{Y}=(I+X-M-sY)$, $\dot{I}=(v\dot{Y}-I)$, and $\dot{X}-\dot{M}=(m\nabla^2 Y-X+M)$. We can again eliminate the variables I, X, and M by differentiating the first equation once more and using the original system for the elimination. In this way we obtain

$$\ddot{Y}+(1+s-v)\dot{Y}+sY-m\nabla^2 Y=0. \tag{2}$$

As demonstrated in Puu (1982) the solution separates for $Y=T(t)\,S(x,y)$, so that

$$T''+(1+s-v)T'+(\lambda m+s)T=0 \quad \text{and} \tag{3}$$

$$\nabla^2 S+\lambda S=0. \tag{4}$$

Obviously (3) is of the same form as the original model (1), so that temporal change for each given λ is that of a damped or antidamped simple harmonic oscillator. The damping coefficient $\alpha = -(1+s-v)/2$ is independent of λ. This is not true about the period $\omega = \sqrt{(\lambda m + s - \alpha^2)}$. We note that the motion is oscillatory whenever $4(\lambda m + s) > (1+s-v)^2$.

From the general theory of eigenvalues and eigenfunctions we, however, know that there exists an infinite spectrum of ascending positive eigenvalues λ and corresponding eigenfunctions $S(x, y)$ for any well defined region of twospace with reasonable boundary conditions. Good references about this are Courant and Hilbert (1953, Vol. 1), or Duff and Naylor (1966).

Due to the superposition principle for linear systems the solution can be written in full generality as

$$Y = \sum_i \exp(\alpha t)\{A_i \cos(\omega_i t) + B_i \sin(\omega_i t)\} S_i(x, y), \tag{5}$$

where S_i are the different eigenfunctions that solve the eigenvalue problem (4) and ω_i are the periods computed from the corresponding eigenvalues λ_i.

Examples of eigenfunctions can be found in Beckmann and Puu (1985) for the rectangular, circular, and spherical regions, i.e., sines, cosines, Bessel functions, and Legendre Polynomials. As all such eigenfunction systems can be chosen so that they form an orthonormal set, it is possible to evaluate the constants A_i and B_i from initial conditions by the integrals

$$A_i = \int\int Y(0) S_i(x, y)\, \mathrm{d}x\, \mathrm{d}y, \tag{6}$$

$$B_i = \int\int Y'(0) S_i(x, y)\, \mathrm{d}x\, \mathrm{d}y, \tag{7}$$

where it is understood that the integration is on all the region in twospace for which the analysis is formulated.

It should be noted that the conclusions are not limited to the exemplified simple types of regions, but hold independent of the shapes. Moreover, as the generalized Sturm–Liouville theory demonstrates, we could even make the propensity to import location-dependent so that we replace $m\nabla^2 Y$ in eq. (2) by the far more general expression $\nabla \cdot (m\nabla Y) = m\nabla^2 Y + \nabla m \cdot \nabla Y$. Even if we thus make the expenditure diffusion by trade dependent on location to account for varying transportation facilities, all the results can be retained.

From (6)–(7) we see that *any periodic change, however irregular its time profile is, can be produced by the spatial multiplier–accelerator model.*

Moreover, as there is no reason for the eigenvalues λ_i and the corresponding periods ω_i to take on rational proportions only, we conclude that *the period of composite cyclical motion may be varying, so that there is no periodicity in a strict sense at all.*

Accordingly, the spatial extension of the simplest business cycle mechanism of all may produce temporal change of any realistic irregularity desired, and this is attained without introducing any distributed lag structure.

However, the model still has one major deficiency, as it produces change that either is bound to go to extinction and stationary quiescence, or tends to increase beyond any limit and blow the whole system up, depending on the sign of α. The only escape is a zero α when the structural coefficients are in the relation $v-s=1$, which would be a very unlikely coincidence.

4. Non-linearity in the original model

This strange property of explosion or extinction is a characteristic of linearity. Linearity, however, in economics as in physics or any other science, can only be a first approximation. Such linearisation is only reasonable when the variables remain within certain bounds. If the model creates change of the variables beyond any limit, then the conditions for linearisation are always violated, and the model becomes self-destructive. This is the absurdity of linear economic models of economic growth or cyclical change. Linearity holds for marginal changes, but the models implicitly claim global relevance as they are used to explain long-run development.

In addition there are strong factual reasons for making the model non-linear, given by Sir John Hicks in his reasoning of a 'floor' to disinvestments when no capital is replaced and is depreciating at its natural rate, and a ceiling to investments when other inputs than capital become binding and their own rate of growth limits investments. Hicks (1950) introduced the constraints as linear inequalities, but the analytical treatment is easier if we incorporate the non-linearity into the investment function. The simplest function, having the properties of a non-linear accelerator with a floor and a ceiling is $v \tanh \dot{Y}$ as a replacement for $v\dot{Y}$. Around zero the function is almost linear in the argument, but for large negative or positive $\dot{Y}$ it goes asymptotically to ± 1.

It should be noted that the numerical values of the asymptotes or the exact form of the function are not meant to have any significance at all. Nevertheless the choice is not one of arbitrary exemplification. With respect to the qualitative behaviour of the model the choice is generic, i.e., it is topologically equivalent to the behaviour of any S-shaped curve having upper and lower asymptotic bounds. In any case where calculations are involved we replace $\tanh \dot{Y}$ by its truncated Taylor series $(\dot{Y}-\frac{1}{3}\dot{Y}^3)$, which has the same qualitative properties as the function itself as regards the behavior of the model.

Let us start by incorporating the non-linearity in the original model, without any interregional trade, as it is formalized in (1). It then becomes

$$\ddot{Y}+(1+s)\dot{Y}-v\tanh\dot{Y}+sY=0. \tag{8}$$

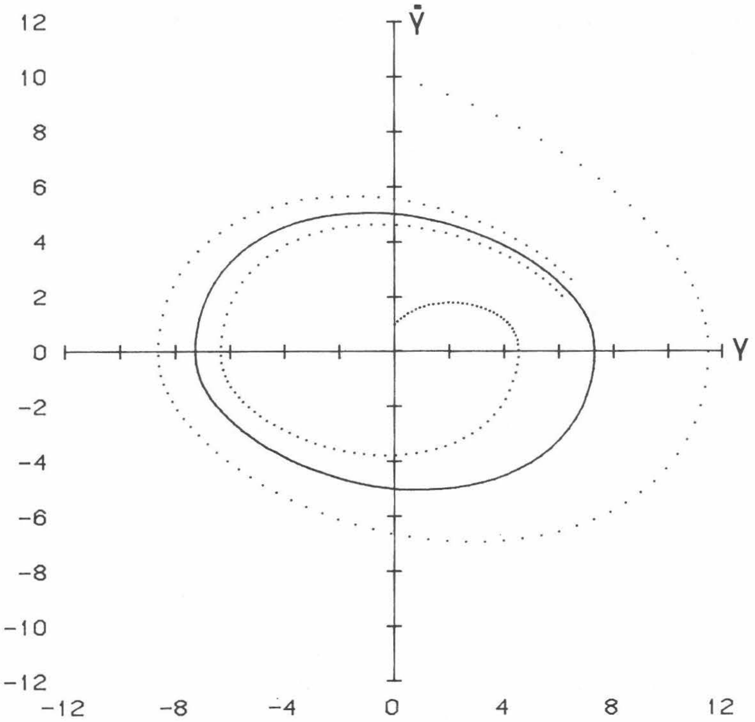


Fig. 1. Stable limit cycle.

The qualitative behaviour of this model can be inferred by sketching a phase diagram in $Y, \dot{Y}$-space. Obviously the system is damped for large $|Y|$ or $|\dot{Y}|$. In the case $v>(1+s)$, there is a neighbourhood of the origin in phase space where there is antidamping, whereas otherwise the damped zone extends to the whole plane. It is the first case that is more interesting. The combination of antidamping in the centre and damping in the periphery makes one suspect the existence of a limit cycle. Actually, numerical integration of (8), whose result is shown in fig. 1, clearly shows the limit cycle.

Another way to proceed is by using the truncated Taylor series written above. Then a change of time scale by the linear factor $\sqrt{s}$ puts the differential equation in the form

$$\ddot{Y}+Y=a\dot{Y}-b\dot{Y}^3, \tag{9}$$

where $a=\sqrt{s}(v-s-1)$ and $b=\sqrt{s^3}/3$ are positive constants. Eq. (9) is very close to Rayleigh's equation for a bowed string. See Jordan and Smith (1977) or Kevorkian and Cole (1981). It is related to the well-known van der Pol oscillator by a simple transformation, and it can be easily approximated by use of one of the singular perturbation methods. The substance of the method is that (9) is regarded as a perturbation of the linear equation $\ddot{Y}+Y=0$, which has a periodic solution of the given period 2π of arbitrary phase and amplitude. One then assumes that the non-linear equation (9) is periodic as well, so that there is a point in separating time in two variables, one fast time representing cyclic motion and one slow time representing any trend factor such as asymptotic approach to a limit cycle. Slow time is defined in relation to the coefficient of the disturbing first derivative in the right-hand side of (9) as $\tau=at$, and one attempts to find a solution of the form $Y=Y^0(\tau,t)+aY^1(\tau,t)+a^2Y^2(\tau,t)+\dots$. Inserting this solution in (9), using the definition of slow time, and assembling terms of order 1, a, a^2 etc. gives a number of relations, the first of which are

$$\ddot{Y}^0+Y^0=0, \tag{10a}$$

$$\ddot{Y}^1+Y^1=-2Y^0_{t\tau}+aY^0_t-b(Y^0_t)^3, \tag{10b}$$

where the subscripts denote differentiation. The solution to (10a) has already been stated:

$$Y^0=A(\tau)\cos t+B(\tau)\sin t, \tag{11}$$

where the assumption that phase and amplitude depend on slow time has been incorporated in the dependence of A and B.

To find out the exact form of the dependence of coefficients we proceed to the next equation, (10b), where the derivatives of (11) are inserted in the right-hand side. Accordingly,

$$\ddot{Y}^1+Y^1=\left(2A'+\frac{3}{4}\frac{b}{a}A(A+B)-A\right)\sin t+\left(2B'+\frac{3}{4}\frac{b}{a}B(A+B)-B\right)\cos t$$

$$-\frac{1}{4}\frac{b}{a}A(A^2-3B^2)\sin 3t-\frac{1}{4}\frac{b}{a}B(B^2-3A^2)\cos 3t,$$

where the primes denote differentiation and where we have used the usual formulas for the cubes of sines and cosines. Next we note that in the integration for Y^1 the $\sin t$ and $\cos t$ terms would give rise to terms of the type $t\cos t$, $t\sin t$ and hence make the series expansion non-uniform. So, we have to set the coefficients of these terms equal to zero. This results in two differential equations for A and B, which seem to be coupled but split if we transform into polar coordinates $A=\rho\cos\phi$ and $B=\rho\sin\phi$. Then,

$$\rho'+\frac{3b}{8a}\rho^3-\tfrac{1}{2}\rho=0 \quad\text{and}\quad \phi'=0.$$

The second equation yields a constant phase angle, whereas the first one may be easily integrated, as it is of the Bernouilli type. So finally

$$Y^0 = c(1-(1-c^2/c_0^2)\exp(-at))^{-1/2}\cos(t-\phi), \tag{12}$$

where $c=\sqrt{(4a/3b)}$ is the amplitude of the limit cycle and c_0 is any initial value. It is seen that there is a uniform approach to the limit cycle $Y^0 = c\cos(t-\phi)$ in this first approximation. The process can be continued by deriving more terms in the power series. The calculations, however, become more and more involved, and the first approximation shows us sufficiently much of the qualitative character of the solution. It need hardly be said that there is good agreement with the result of numerical solution.

5. Bifurcations

We can see by inserting the definitions of a and b that $c=2\sqrt{((v-s-1)/s)}$. Accordingly, the amplitude of the limit cycle increases with v and decreases with s. The dependence is continuous in character until the critical relation $v=s+1$ is fulfilled. For higher s or lower v the limit cycle no longer exists. The system is globally damped, as was seen above. This phenomenon is an example of a bifurcation.

More interesting examples may be obtained with a more complicated investment function, such as for instance $\tanh^3 \dot{Y}$. This may be approximated by $\dot{Y}^3 - \dot{Y}^5$. Such an investment function results in the following phenomena. Around the origin there is damping until an unstable limit cycle is reached. Outside this there is a zone with antidamping, while there is damping again outside an outer stable limit cycle. This is shown in fig. 2.

For the approximation $\dot{Y}^3 - \dot{Y}^5$ we deal with the equation $\ddot{Y} + sY = v(\dot{Y}^3 - \dot{Y}^5) - (1+s)\dot{Y}$, which after a linear change of time scale with the factor $\sqrt{s}$ takes the form

$$\ddot{Y} + Y = -(\sqrt{s} + 1/\sqrt{s})\dot{Y} + v\sqrt{s}\dot{Y}^3 - v\sqrt{s^3}\dot{Y}^5. \tag{13}$$

Using the same singular perturbation method as in the previous case, we obtain the first approximation

$$Y^0 = A(\tau)\cos t + B(\tau)\sin t. \tag{14}$$

In order that no secular terms of the type t cos or $t\sin t$ should turn up in the solution of the next approximation Y^1 we again get a constraint on A and B in terms of a pair of differential equations. By introducing polar coordinates as before we arrive at the system

$$\rho' + \tfrac{5}{16}v\sqrt{s^3}\rho^5 - \tfrac{3}{8}v\sqrt{s}\rho^3 + \tfrac{1}{2}(\sqrt{s} + 1/\sqrt{s})\rho = 0 \quad \text{and} \quad \phi' = 0,$$

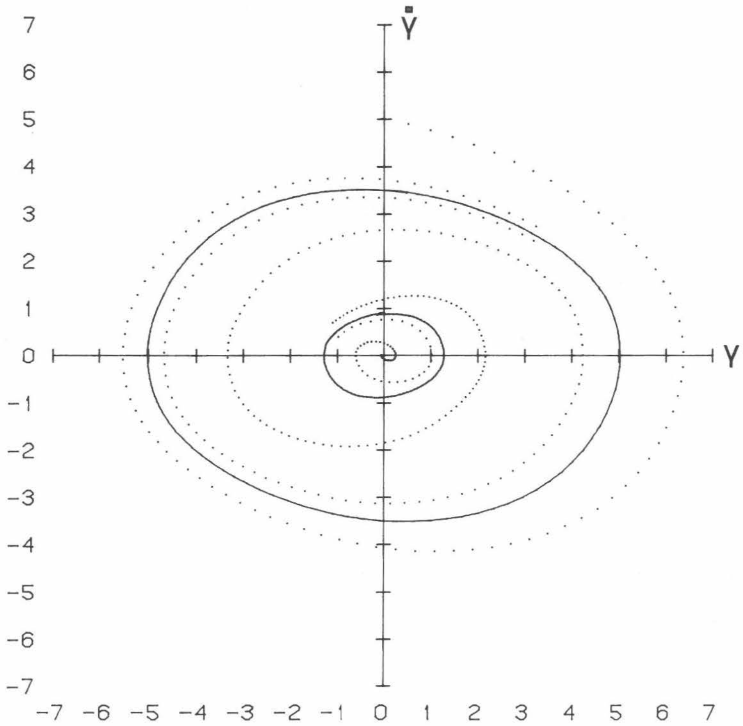


Fig. 2. Stable and unstable limit cycles.

equivalent to the pair of equations above. The second one again yields a constant phase angle, but the first is not as readily integrated as before. Therefore, let us introduce the change of variable $z = s\rho^2$, so that

$$\sqrt{s}z' = -(1+s)z + 3/4vz^2 - 5/8vz^3. \tag{15}$$

By interchanging dependent and independent variables this can be written

$$-(\sqrt{s} + 1/\sqrt{s})\mathrm{d}\tau = \mathrm{d}z \Big/ \left(z - \frac{3}{4}\frac{v}{1+s}z^2 + \frac{5}{8}\frac{v}{1+s}z^3 \right)$$

$$= \left(\frac{1}{z} + \frac{\gamma}{\alpha - z} + \frac{\delta}{\beta - z} \right) \mathrm{d}z. \tag{16}$$

This last splitting in partial fractions works provided we choose

$$\alpha, \beta = 3/5(1 \pm \sqrt{(1-40/9(1+s)/v)}) \quad \text{and}$$

$$\gamma, \delta = 1/2(1 \pm 1/\sqrt{(1-40/9(1+s)/v)}).$$

With this choice the differential equation can be integrated to

$$k\exp(-(\sqrt{s}+1/\sqrt{s})\tau) = \lambda = z(\alpha - z)^{\gamma}(\beta - z)^{\delta}.$$

As an example let us choose $v = 5(1+s)$. Then $\alpha = 0.8$, $\beta = 0.4$, $\gamma = 2$, and $\delta = -1$. Accordingly

$$\lambda = \frac{z(0.8-z)^2}{(0.4-z)},$$

with $\lambda = k\exp(-4\tau)$ if we choose $(1+s) = 4\sqrt{s}$, i.e., $s = 0.07$. Inverting the preceding equation means solving a third degree equation

$$z^3 - 8/5z^2 + ((4/5)^2 + \lambda)z - 2/5\lambda = 0.$$

Defining $Q = -(4/15)^2 + \lambda/3$ and $R = -(4/15)^3 - \lambda/15$ we obtain the discriminant

$$D = Q^3 + R^2 = \tfrac{3}{2}(\tfrac{4}{15})^4\lambda - \tfrac{13}{48}(\tfrac{4}{15})^2\lambda^2 + \tfrac{1}{27}\lambda^3.$$

Factoring λ out leaves a positive definite quadratic form. Accordingly, the discriminant is positive or negative depending on the sign of λ. With the discriminant positive there is only one real root given by

$$z = \tfrac{8}{15} + \sqrt[3]{(R+\sqrt{D})} + \sqrt[3]{(R-\sqrt{D})}.$$

With a negative λ there are three real roots, one of which is necessarily negative and hence of no interest. The positive roots then are

$$z = \tfrac{8}{15} + \sqrt{-Q}(\cos\tfrac{1}{3}\cos^{-1}(-R/\sqrt{-Q^3}) \pm \sqrt{3}\sin\tfrac{1}{3}\cos^{-1}(-R/\sqrt{-Q^3})).$$

To find out the asymptotic behaviour we note that the limiting value of λ is zero as τ increases without any bound. The limiting values are $Q = -(4/15)^2$, $R = -(4/15)^3$ and $D = 0$. Accordingly the limiting values of z are 0 and 0.8.

Thus the origin and a limit cycle with amplitude 0.8 are the attractors of the system. The critical value $\beta = 0.4$ of z, on the other hand, represents an unstable limit cycle. There is no bifurcation in the case illustrated. However, we see from our considerations that for $9v=40(1+s)$ we have $\alpha=\beta=0.6$, so that the limit cycles fuse. Accordingly, the stable limit cycle of finite amplitude vanishes at this combination of structural coefficients.

Accordingly, provided the parameters s and v undergo continuous change so that such a coincidence of limit cycles occurs, then the system that has formerly been in motion along a limit cycle of finite amplitude suddenly collapses to stationary equilibrium.

6. Non-linearity in the spatial model

Let us now turn to a much tougher task, namely to introduce the non-linear investment function $v \tanh \dot{Y}$ into the spatial business cycle model (2). Little is known about non-linear partial differential equations, and what is known points at the possibilities of a wide variety of solutions some of which are manifestly strange. Anyhow, the differential equation we deal with is

$$\ddot{Y}+(1+s)\dot{Y}-v\tanh \dot{Y}+sY-m\nabla^2 Y=0. \tag{17}$$

We can immediately see two possible solutions. One is to put $\nabla^2 Y\equiv 0$, so that the system becomes identical with (8) above. Accordingly, the limit cycle solution with income at all locations changing in phase is a possibility. Another solution is to put $\ddot{Y}=\dot{Y}=0$ and require that the purely spatial equation $sY=m\nabla Y$ is fulfilled.

Accordingly, there seems to be a tradeoff between temporal change and spatial inhomogeneity. Suppose that we deal with a bounded region where $\dot{Y}=0$ on the boundary. Then we multiply (17) by $\dot{Y}$, integrate over the whole region and use Gauss's integral theorem. So we obtain

$$\mathrm{d}/\mathrm{d}t \iint \dot{Y}^2+sY^2+m(\nabla Y)^2\,\mathrm{d}x\,\mathrm{d}y=2\iint v\dot{Y}\tanh \dot{Y}-(1+s)\dot{Y}^2\,\mathrm{d}x\,\mathrm{d}y.$$

The left-hand side represents the temporal rate of change of the 'energy' of the system, as composed by temporal change and/or spatial inhomogeneity 'tension'.

The right-hand side is positive for small $|\dot{Y}|$ and negative for large $|\dot{Y}|$. Accordingly, the system gains energy when $\dot{Y}$ is predominantly small. Suppose that movement is close to extinction. Then the system must be gaining energy either in terms of increasing $|\dot{Y}|$ again, or in terms of spatially increasing tensional inhomogeneity. Movement can be frozen down in increased spatial inhomogeneity, but otherwise it can never die out, provided $v>(1+s)$. This condition is required for our reasoning about the right-hand side and it is equivalent to the condition for existence of a limit cycle in the model in section 4.

If we regard the terms involving $\dot{Y}$ as a perturbation we see that what remains is a two-dimensional variant of the linear Klein–Gordon equation. See Whitman (1974) or Nayfeh (1973). Its solution is known to produce dispersive waves. So, let us try a solution of the form $Y(\theta)$ with $\theta = k\sqrt{(x^2+y^2)} - \omega t$. Provided the dispersion relation $mk^2 - \omega^2 + s = 0$ is fulfilled the linear equation is solved. For definiteness and simplicity of calculation put $m = s = 0.5$. Accordingly, the dispersion relation allows a wave number $k = 1$ along with a period $\omega = 1$. Moreover, we put $v = 2$ and denote the spatial radius vector by $r = \sqrt{(x^2+y^2)}$. This transforms (17) into

$$Y'' + Y = 4\tanh Y' - (3 + 1/r)Y', \tag{18}$$

where Y' and Y'' denote the derivatives of the assumed solution with respect to the argument $\theta = r - t$.

This equation is similar to (8), except for the spatial dependence on r, and may be dealt with by similar methods as were used for (8). Fig. 3 illustrates the limit cycles for various values of r. For radii less than unity the system is damped and there exist no limit cycles. For larger radii limit cycles with finite and increasing amplitude emerge, the outmost one corresponding to an

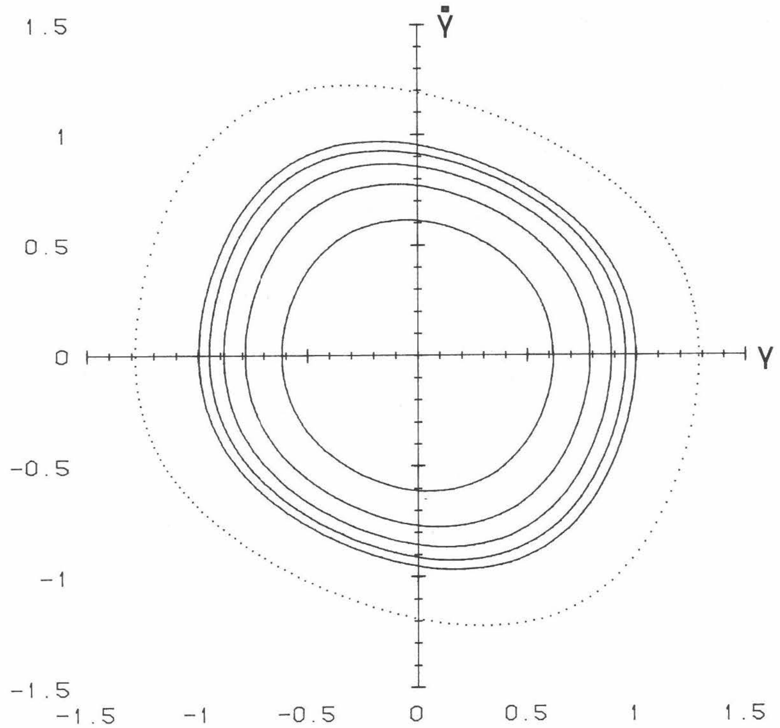


Fig. 3. Limit cycles at different distances from the origin of x, y space, $r = 1.5, 2, 2.5, 3, 3.5$.

infinite radius. Although the amplitudes of the cycles increase with distance from the origin of x, y space, the periods seem to be the same. This is confirmed by direct calculation according to a method used in Jordan and Smith (1977).

A first approximation to the linear left-hand side of (18) put equal to zero is a harmonic motion of period 2π and amplitude 1. So, $Y = \sin\theta$ is a solution. Next, we insert the solution in the right-hand side of (18) multiply by Y', and integrate over a whole cycle to get the next approximation

$$T = 2\pi + 4\int_0^{2\pi} \tanh(\sin\theta)\cos\theta\,\mathrm{d}\theta - (3 + 1/r)\int_0^{2\pi} \sin\theta\cos\theta\,\mathrm{d}\theta.$$

But, the hyperbolic tangent being symmetric around zero and of the same sign as its argument, both integrals vanish. We see that not only is the period T independent of r up to next approximation, but the original period of the linear equation remains unchanged. This observation confirms the results of numerical solution.

Finally, we can approximate $\tanh Y'$ by $Y' - Y'^3/3$, so that (18) becomes

$$Y'' + Y = (1 - 1/r)Y' - \tfrac{4}{3}Y'^3. \tag{18a}$$

This equation may be treated with the same two-time method as was used for (9). The final solution for the first term in the series is

$$Y^0 = (1 - 1/r)((1 - (1 - 1/r)^2/c_0^2)\exp((1 - 1/r)(r - t)))^{-1/2}\cos(r - t).$$

This establishes the limit cycles of an amplitude that increases with r. The system obviously approaches a limit cycle provided that r is at least unity, otherwise we see that the damping does not function. We also see that the approach to the limit cycle is very slow when r is close to the critical unitary value.

The asymptotic solution to the first approximation is

$$Y^0 = (1 - 1/r)\cos(r - t).$$

For any given moment t this gives a spatial pattern. It is illustrated by a computer diagram in fig. 4. It is interesting to note that in this case the spatial coordinate acts as a parameter that causes bifurcation in temporal behaviour. It must however be remembered that we have investigated only a few of the phenomena that the non-linear spatial model can produce.

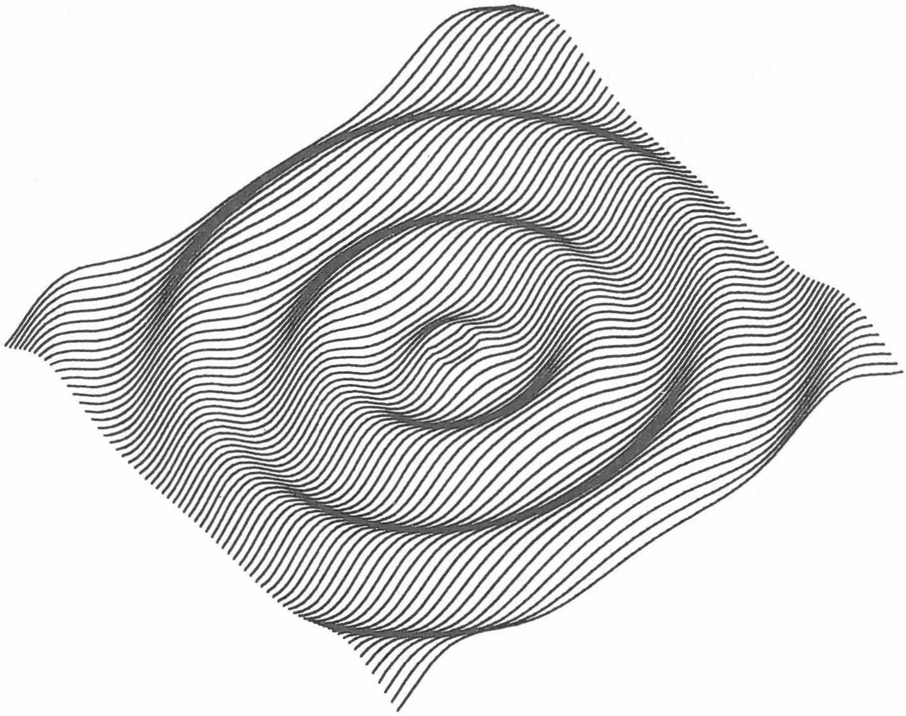


Fig. 4. Spatial pattern in the asymptotic solution.

7. Conclusion

We have demonstrated that two features of the original multiplier–accelerator model of the business cycle: the regularity of its solution and the tendency to explode or extinguish any dynamics in the system can be removed by putting the model in a spatial context and by making it non-linear.

References

Allen, R.G.D., 1956, Mathematical economics (Macmillan, New York).
Beckman, M.J. and T. Puu, 1985, Spatial economics (North-Holland, Amsterdam).
Courant, R. and D. Hilbert, 1953, Methods of mathematical physics (Wiley-Interscience, New York).
Duff, G.F.D. and D. Naylor, 1966, Differential equations of applied mathematics (Wiley, New York).
Goodwin, R.M., 1951, The non-linear accelerator and the persistance of business cycles, Econometrica 19, 1–17.
Hicks, J.R. (1950), A contribution to the theory of the trade cycle (Oxford University Press, Oxford).

Jordan, D.W. and P. Smith, 1977, Non-linear ordinary differential equations (Clarendon Press, Oxford).
Kevorkian, J. and J.D. Cole, 1981, Perturbation methods in applied mathematics (Springer, New York).
Nayfeh, A.H., 1973, Perturbation methods (Wiley, New York).
Phillips, A.W., 1954, Stabilisation policy in a closed economy, Economic Journal 64, 290–323.
Puu, T., 1982, Outline of a trade cycle model in continuous space and time, Geographical Analysis 14, 1–9.
Samuelson, P.A., 1939, Interactions between the multiplier theory and the principle of acceleration, Review of Economic Statistics 21, 75–78.
Whitham, G.B., 1974, Linear and non-linear waves (Wiley-Interscience, New York).

Regional Science and Urban Economics 16 (1986) 97–122. North-Holland

ON OCCUPATION STRUCTURE AND LOCATION PATTERN IN THE STOCKHOLM REGION*

Christer ANDERSTIG and Björn HÅRSMAN

Regional Planning Office, Stockholm, Sweden

Received October 1984, final version received September 1985

An entropy model is used to decompose employment changes into effects of industry, occupation and region. The results from an application to Sweden indicate that metropolitan areas in general, and Stockholm in particular, have locational advantages in knowledge-oriented activities. The location of employment and firms in the Stockholm region is analyzed by using both an entropy model and a logit model. The results demonstrate the importance of introducing different measures of accessibility. The location of knowledge-oriented activities is shown to be significantly influenced by the location pattern of the labour force.

1. Introduction

This paper is mainly explorative. It reports from the initial work undertaken in a project recently started at the Stockholm Regional Planning Office. The aim of this project is to develop empirically tractable and behaviourally oriented models that can be used to estimate the impact of infrastructural investments on residential location and workplace location.

There are several reasons for our interest in occupation structure and location pattern. One arises from the socio-economic imbalance between the northern and the southern part of the Stockholm region. To decrease this imbalance has since long been a major policy goal and it has recently been suggested that infrastructural investments in the southern part of the region could be efficient instruments for achieving this end. A second and related reason is the hypothesis that different occupation categories dominate in different phases of product cycles. Hence, the occupation structure of the labour supply may be of importance in the location decisions of firms.

The work reported in the paper is focused on employment and workplace location. Residential location is mainly treated by describing current trends in the pattern of residential location for different occupation groups. The paper is organized as follows. Section 2 gives a presentation of the economic structure in the Stockholm region as compared to other Swedish regions. Section 3 follows up by applying an entropy model in decomposing employment (−changes) into effects of industry, occupation and region. In

*The authors benefited from comments by an anonymous referee. The project is supported by a grant from the Swedish Council for Building Research.

section 4 the same model is used for intraregional analyses with respect to the Stockholm region. Applying the entropy models, the resulting dual variables are the basis of discussing the importance of occupation for workplace and residential location. As regards intraregional workplace location, this issue is further discussed in section 5. Applying some logit models, attempts are made to catch the effects of infrastructural investments on firm location.

2. The occupation structure in an interregional perspective

Like in all developed countries the employment structure in Sweden has during last decades been changing from manufacturing towards services, a trend that can be shown in terms of industry and occupation, as in fig. 1. In general terms this structural change is due to a rapid productivity growth in manufacturing, by which this industry has grown less manual, and to a high elasticity of demand for publicly – and privately – provided services.

Although the structural change shown in fig. 1 represents a general tendency for all regions, there is a persisting pattern of interregional

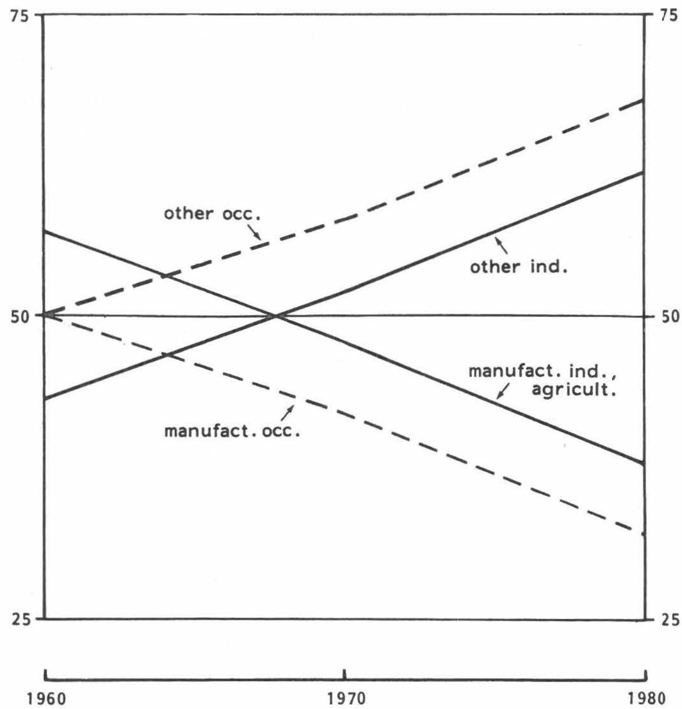


Fig. 1. Total employment in Sweden by two industry groups and two occupation groups 1960–1980 (percent).

differences as regards industry and occupation orientation. Analyses of regional hierarchies of occupations have been performed by, e.g., Burns and Healy (1978) presenting an economic interpretation of Central Place Theory. Its principal tenet, that the existence of scale economies produces a spatial distribution of economic activities varying with the spatial distribution of population, is developed in a paper by Andersson and Johansson (1984b). Their interest concerns the distribution of knowledge-oriented activities in particular, as they use the theory of product cycles to explain regional specialization.

In short the theory states that in different phases in the product cycle the centre of production and employment is shifting interregionally as the product develops from a high R&D-intensity to maturity. It can be envisaged that the metropolitan areas have a locational competitive advantage in the first phases of the cycle, dominated by research, knowledge-oriented activities and technical and commercial development. The idea that the Stockholm region, and other metropolitan areas, have competitive advantages in these respects could be founded on a lot of historical and institutional circumstances.

In comprehension, a high level of accessibility is a general basis of explanation. By reasons of indivisibles and the existence of minimum market sizes, knowledge-oriented activities tend to concentrate to metropolitan areas and other areas with a high level of intraregional accessibility. This means, on the other hand, that for activities with a lower dependence on intraregional accessibility, and hence lower bid-prices for central land, the metropolitan areas will have a locational competitive disadvantage.

In the following we are illustrating regional specialization in an attempt to classify regions according to the level of intraregional accessibility. All Swedish municipalities are grouped into seven 'H-regions'. Although the definitions refer to the situation in 1980 the regions are the same during the 20 years covered by the study.

- H1 Stockholm.
- H2–H3 The other two metropolitan areas (H2: Gothenburg and H3: Malmö).
- H4 Municipalities with more than 90 000 inhabitants within their central labour market area, 30 km.
- H5 Municipalities with 27 000–90 000 inhabitants as described above and with more than 300 000 inhabitants within a radius of 100 km from the municipal centre.
- H6 Municipalities with 27 000–90 000 inhabitants within 30 km from municipal centre and with less than 300 000 within 100 km from the municipal centre.
- H7 Other municipalities (rural municipalities with less than 27 000 inhabitants within a radius of 30 km from the municipal centre).

In fig. 2 the economic structure is represented by five occupation groups which try to characterize different types of work according to the level of knowledge orientation, somewhat ad hoc defined. *Technical, scientific work* includes physicists, chemists, engineers, architects, physicians, biologists etc.; *Other knowledge based work* includes social scientists, teachers, programmers, journalists, lawyers, judges, artists etc.; *Administrative work* includes managers, accountants, cashiers, secretaries, typists etc.; *Service-oriented work* includes nurses, clergymen, tradesmen, busdrivers, policemen, cooks, barbers etc. and finally *Manufacturing, agricultural work* includes farmers, fishermen, craftsmen and all kind of workmen in manufacturing, construction etc. The main reason for combining manufacturing and agricultural employment is that both groups are characterized by goods handling work. Another is convenience; in Sweden agricultural work corresponded to around ten percent of total employment in 1960 and five percent in 1980.

On the horizontal axis the regional share of total national employment is shown and on the vertical axis the occupational share of total regional employment. Thus the relative size of regional employment is shown by the corresponding areas in the figure.

The figure illustrates that all types of work, except service-oriented, prove to be almost monotonically increasing or decreasing functions of intra-regional accessibility. In the Stockholm region Technical, scientific work, and Other knowledge based work represent 15 percent of regional employment in 1960 and 23 percent in 1980.

The fact that region H6, and to some extent region H3, diverge from the expected pattern is worth a comment. Region H6 is dominated by some regional centres in northern Sweden. These have since the sixties grown into centres of higher education and R&D, in main by location of new universities and supporting infrastructure, such as airports etc. By the same token, the predominant role played by the university in region H3 is a distinctive character as compared to region H2, although both regions are centres of higher education and R&D.

The question of possible impact of regional centres of R&D upon the economic structure in general is for one thing a reason to combine occupation data with industry data. Fig. 3 presents the intensity of knowledge-oriented occupations, i.e., the sum of Technical, scientific work and Other knowledge based work, in regional employment for three aggregates of industries.

By differentiating between private and public industries the structural differences between regions become even more apparent.

In manufacturing and private services the knowledge-intensity in the metropolitan areas, particularly Stockholm, is much higher than in other regions. For the four less dense populated regions, H4–H7, the knowledge-intensity in year 1980 has approached that of the Stockholm region in year

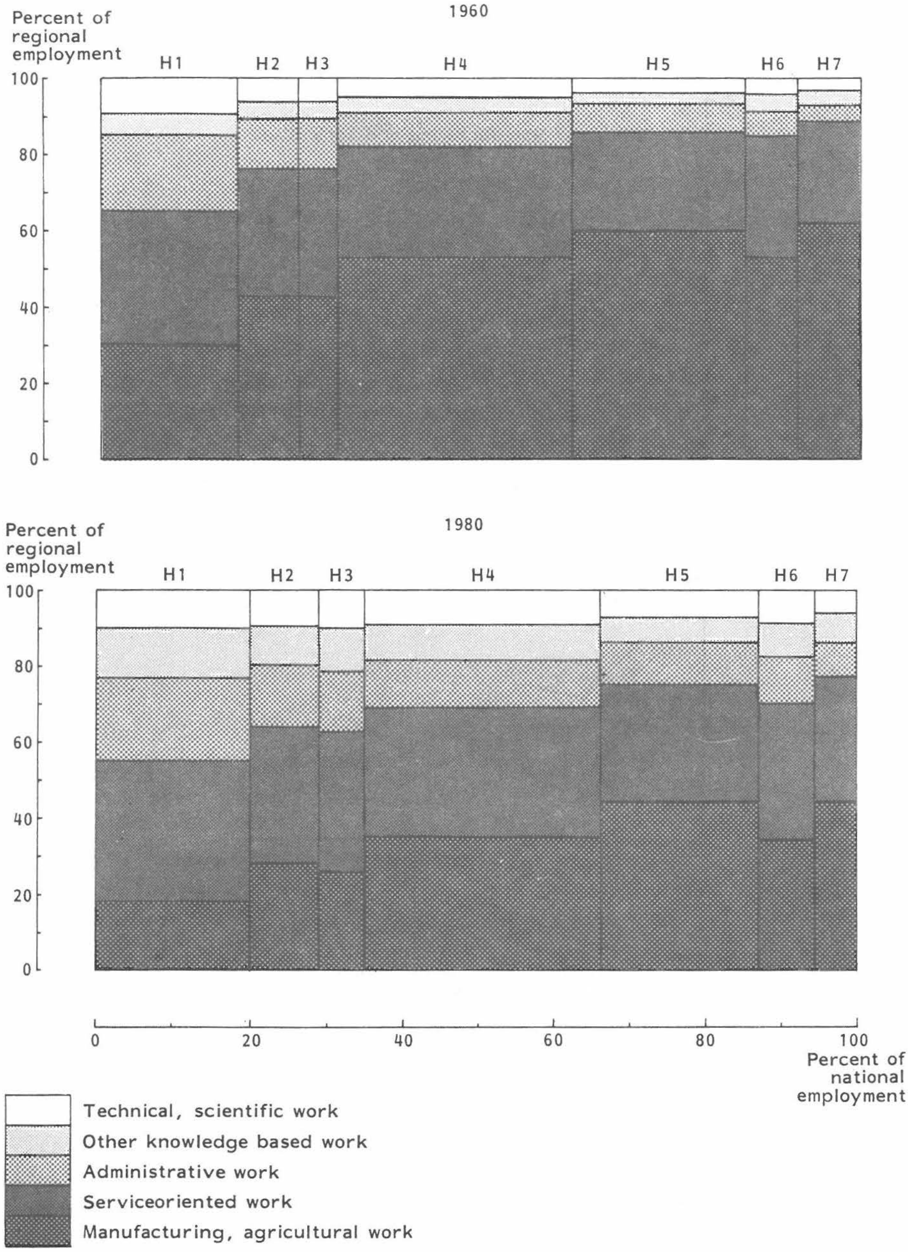


Fig. 2. Regional employment (20-w hours/week) by occupation (type of work) in seven H-regions in 1960 and 1980 (percent).[a]

[a]Employed persons with unspecified location are not included.

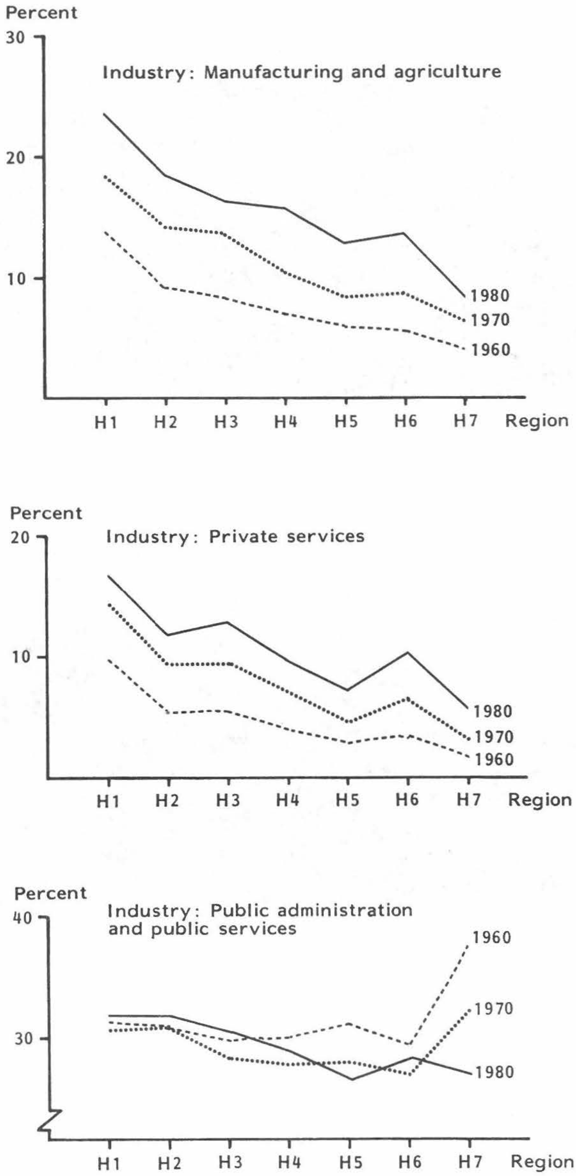


Fig. 3. Intensity of knowledge-oriented occupations in three industries in seven H-regions in 1960, 1970 and 1980 (percent).

1960. Thus, in this respect the non-metropolitan areas have a knowledge-intensity lag of twenty years. There are reasons to assume that the future expansion in manufacturing will be concentrated to high-technology sectors. Assuming that accessibility to high-educated personnel is an important interregional location factor we can anticipate a further concentration of high-technology enterprises to the metropolitan areas in general and to the Stockholm region in particular. The assumption of a competitive advantage for the Stockholm region seems even more justified allowing for the influence of the neighbouring R&D-centre in Uppsala, north of Stockholm.

It is interesting to note, however, that region H6 shows relatively high and rapidly rising knowledge intensities in the private industries. This could indicate that the regional centres in question have succeeded in attaching knowledge-intensive activities to the basic knowledge capital, invested in these centres by locating universities.

As to public administration and public services, the knowledge intensity is generally higher and the interregional differences are smaller. The conspicuous decrease for some regions is mainly an effect of industry-mix. E.g., in 1960 education makes nearly 50 percent of sectoral employment in the sparsely populated region H7, compared to 26 percent in Stockholm. During the period 1960–1980 public employment has expanded, especially in non-metropolitan regions, and the mix of public sectors has tended to equalize.

To gain further insight into the question of the spatial distribution of occupations, we shall in the following decompose regional employment in terms of both industry- and occupation-mix.

3. Application of an entropy model to decompose regional employment

Shift-share has been a standard technique in descriptions of regional employment growth, where growth is decomposed into an industrial mix effect and a regional 'competitive' effect. The theoretical underpinning and analytical properties of shift-share analysis have often been questioned. As to the empirical limitations it should be evident from above that the industrial mix is only a rough representation of economic structure.

The emerging pattern from traditional shift-share analysis has generally shown that larger, metropolitan areas are characterized by a positive industrial mix effect (favourable composition of fast-growth industries) and a negative regional competitive effect (slowly growing in the region), whereas the reverse pattern applies to peripheral regions, see e.g., Perloff and Wingo (1961), Paraskevopoulos (1974). This kind of empirical regularity also applies to the regional development in Sweden, see Anderstig (1981).

If we confine ourselves to manufacturing, explanations within the traditional, static Heckscher–Ohlin framework will not be very successful. The relative abundance of knowledge-intensive labour force in the Stockholm region should, according to this theory, imply a specialization in the corresponding, expanding manufacturing sectors. Contradictory, the Stockholm region has shown weak growth or even contraction in these sectors, e.g., electronics, during the last decade.

An interesting approach is presented by Norton and Rees (1979), taking the product cycle model as a major explanatory mechanism behind the process of manufacturing decentralization in the U.S. Their point, however, is not to explain the traditional regional roles. Using shift-share analysis for empirical tests, they suggest that the core regions, the industrial heartland, has lost its sustaining seedbed role and that the high technology growth sectors are manifesting themselves in the growing peripheral states of the American South and West.

We think that this kind of analysis could be elaborated, both with respect to basic data and the model applied. First, the product cycle theory can essentially be seen as a dynamic version of the factor proportions theory. As we have argued above the product cycle theory proposes that the core regions (metropolitan areas) should have a locational competitive advantage in the first *phases* of the product cycle. This proposition does not imply that metropolitan areas should specialize in fast-growing *sectors* as such. On the assumption that innovative capacity can be represented by the knowledge-intensity of employment, the 'seedbed' role should thus be judged both by the occupation and industrial mix, not only the latter.

Second, an entropy model is to prefer to shift-share analysis. As argued by Theil and Gosh (1980) the entropy model has several advantages, inter alia a justification based on concepts from information theory. As shown by Snickars and Weibull (1977) an entropy model produces the most probable estimates, given the information of a priori values and constraints. Further, the entropy model makes it quite simple to include interaction effects, e.g., the interaction between occupation and region.

Let Q^t_{ijk} denote the observed and P^t_{ijk} the estimated number of employed in industry i and occupation group j in region k. We write

$Q^t_{i..} = \sum_j \sum_k Q^t_{ijk}$ total number of employed in industry i in year t,

$Q^t_{.j.} = \sum_i \sum_k Q^t_{ijk}$ total number of employed in occupation group j in year t,

$Q^t_{..k} = \sum_i \sum_j Q^t_{ijk}$ total number of employed in region k in year t,

$Q^t_{ij.} = \sum_k Q^t_{ijk}$ total number of employed in industry i and occupation group j in year t,

$Q^t_{.jk} = \sum_i Q^t_{ijk}$ total number of employed in occupation group j in region k in year t,

$Q^t_{i.k} = \sum_j Q^t_{ijk}$ total number of employed in industry i in region k in year t.

The model is applied to the seven H-regions and five occupation groups already defined, and to the following twelve industries: (1) Agriculture and forestry, (2) Heavy industry, (3) Light industry, (4) Engineering, (5) Construction, (6) Local services, (7) National and regional services, (8) Transportation, (9) Public administration, (10) Health services, (11) Education and research, (12) Social welfare.

The model is given by eqs. (1)–(8). As will be commented on below constraints (3)–(8) have been included after correcting for (2).

$$\operatorname{Min} \sum_i \sum_j \sum_k P^t_{ijk} \ln (P^t_{ijk}/Q^{t-1}_{ijk}), \quad \text{subject to} \tag{1}$$

$$\sum_i \sum_j \sum_k P^t_{ijk} = Q^t_{...}, \tag{2}$$

$$\sum_j \sum_k P^t_{ijk} = Q^t_{i..}, \qquad \sum_i \sum_k P^t_{ijk} = Q^t_{.j.}, \qquad \sum_i \sum_j P^t_{ijk} = Q^t_{..k}, \qquad (3)(4)(5)$$

$$\sum_i P^t_{ijk} = Q^t_{.jk}, \qquad \sum_j P^t_{ijk} = Q^t_{i.k}, \qquad \sum_k P^t_{ijk} = Q^t_{ij.}. \qquad (6)(7)(8)$$

If (2) were the only constraint to be included the estimated values would be the a priori values, Q^{t-1}_{ijk}, multiplied by the factor $Q^t_{...}/Q^{t-1}_{...}$. This multiplier is called national share effect. By correcting for this effect before applying constraints (3)–(8), inclusion of the latter gives information of how the employment change can be attributed to industries, occupations, regions and interaction effects.

If we let λ denote the dual variables associated to constraints (2)–(8) we have

$$P^t_{ijk} = Q^{t-1}_{ijk} \exp [\lambda + \lambda_i + \lambda_j + \lambda_k + \lambda_{ik} + \lambda_{jk} + \lambda_{ij}], \tag{9}$$

i.e., estimated employment changes expressed as a function of the dual variables.

As already indicated λ corresponds to the total national employment growth ($\exp \lambda = Q^t_{...}/Q^{t-1}_{...}$). The coefficients λ_i, λ_j and λ_k indicate how the specific industry, occupation and region has developed relative to other industries, occupations and regions. For example, a positive (negative) value of λ_j means that this occupation has grown (declined) relative to other occupations. The coefficients λ_{ik}, λ_{jk} and λ_{ij} indicate the relative strength of interaction region–industry, region–occupation and industry–occupation, respectively. For example, a positive value of λ_{jk} indicates that occupation j has grown faster in region k than expected from regional growth, λ_k, and occupation growth, λ_j, in general.

As specified in (1) the entropy model will be used to analyze the development 1960–1970 and 1970–1980, with the employment figures of 1960 and 1970 as respective a priori values. To analyze the employment structure of the initial year 1960 we will also use the following criterion function:

$$\mathrm{Min} \sum_i \sum_j \sum_k P^t_{ijk} \ln P^t_{ijk}, \quad \text{subject to} \quad (2)\text{–}(8). \tag{10}$$

It should be emphasized that (10) produces estimates of employment totals whereas (9) provides estimates of employment changes. As a consequence the dual variables corresponding to (10) indicate the economic structure for a specific year. In this case λ indicates a uniform distribution of employment over industries, occupations and regions ($\exp \lambda = Q^t_{...}/I \cdot J \cdot K$, where I, J, K denote number of industries, occupation groups and regions, respectively).

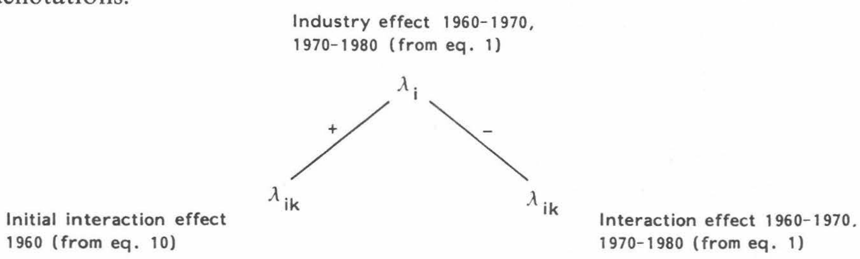
Table 1 presents the regional employment growth during the sixties and the seventies in terms of what here are labeled as regional competitive effects, λ_k. The national share effect, λ, was 1.03 in 1960–1970 and 1.11 in 1970–1980. In this section the effects are represented by the exponential of respective λ:s.

Table 1 indicates that during the sixties the employment growth was concentrated to the metropolitan areas, whereas during the seventies the growth was more decentralized. Again, it is interesting to note the rapid growth in region H6 during the seventies.

Table 1
Employment growth 1960–1970 and 1970–1980 attributed to regional competitive effect.

	Regional competitive effect	
Region	1960–1970	1970–1980
H1 (Stockholm)	1.08	1.01
H2 (Gothenburg)	1.08	1.00
H3 (Malmö)	1.09	0.97
H4	1.03	1.00
H5	0.96	0.96
H6	0.96	1.09
H7	0.72	1.04

According to the discussion about traditional regional roles we should expect the Stockholm region to show a positive industrial mix effect and negative interaction effects region–industry in fast-growing industries, in denotations:



In all essentials, our results come up to the expected pattern. The initial interaction effects are positive for all fast-growing industries (except one) whereas the interaction effects indicating regional development in corresponding industries all are negative. The reverse pattern applies to peripheral regions.

As for the favourable mix in Stockholm, it may be reflected in terms of occupation groups as well. The initial, structural, interaction effects between region and occupation group, the resulting λ_{jk} from (10), show a favourable occupational mix in Stockholm, as the interaction effects are positive for all fast-growing occupations: Technical, scientific work, Other knowledge based work and Administrative work. As regards Other knowledge based work this favourable mix has become even more pronounced since 1960.

Turning to the slowly growing Manufacturing, agricultural work, the corresponding interaction effects are shown in fig. 4.

Fig. 4. Interaction effects between region and Manufacturing, agricultural work in 1960, 1960–1970 and 1970–1980 in seven H-regions.

The initial, structural, interaction effects in 1960 show the regional employment in this occupation group as compared to the expected based on total regional employment and total employment in the occupation group. Obviously there is a hierarchical pattern, such that metropolitan areas, H1–H3, have relatively low initial interaction effects, whereas the non-metropolitan areas, H4–H7, have relatively high initial interaction effects. This pattern also applies to the development 1960–1970 and 1970–1980, although the interaction effect decreases in all regions.

As to the generally fast-growing occupation groups Technical, scientific work and Administrative work, these have grown slower than expected in the Stockholm region. Here we have a special interest as regards the regional interaction effects in Technical, scientific work, shown in fig. 5.

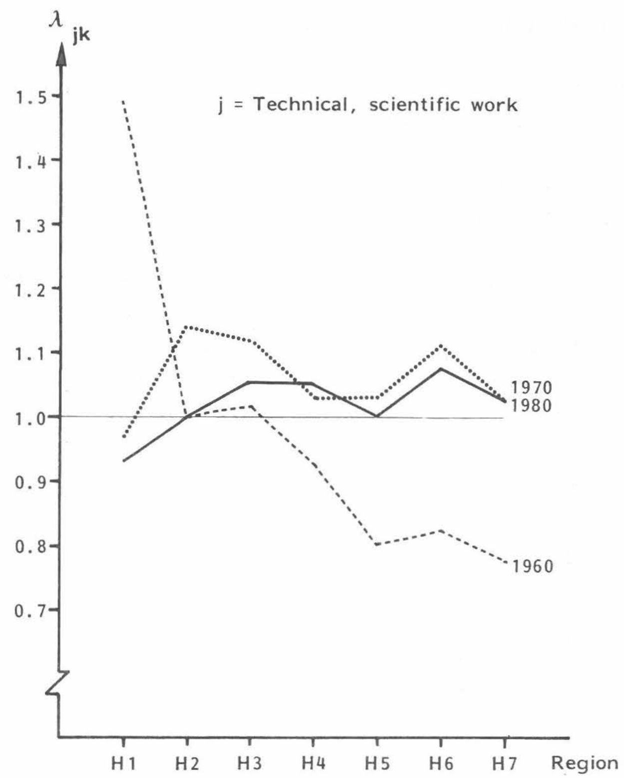


Fig. 5. Interaction effects between region and Technical, scientific work in 1960, 1960–1970 and 1970–1980 in seven H-regions.

The pattern of the initial interaction effects, indicating the structure in 1960, is not surprisingly the reverse of the previous, yet with more pronounced differences between regions.

In 1960 actual employment in Technical, scientific work in Stockholm exceeds the expected number by nearly 50 percent. The development during the sixties and the seventies indicates, however, that this interaction effect has decreased in Stockholm while increased in other regions. This result seems contradictory to our expectations since Technical, scientific work represents R&D-activities better than other occupation groups and, according to the discussion in section 2, there are reasons to assume a strengthened locational competitive advantage in Stockholm for this kind of activities. Even if we a priori have no reasons to qualify this assumption, as to different types of R&D-activities, we must still observe the difference between private and public R&D-activities.

In section 2 we noted that relatively high knowledge intensities in region H6 partly could be explained by the location of universities. In fact

Education and research is the most important industry to explain the relatively high interaction effect for region H6, as illustrated in fig. 5. But location policy as regards public R&D-oriented activities has not been limited only to universities. During the last decades relocations from the Stockholm region of national research institutes, offices of landsurveying, meteorology etc. have been significant elements in regional policy. Interpreting the pattern in fig. 5 we should therefore make a distinction between interaction effects in private and public sectors respectively which can be done by comparing estimated, P^t_{ijk}, and actual, Q^t_{ijk}, values.

Let us first maintain, judging from the general interaction effect between occupation group and industry, λ_{ij}, that five industries have high intensities in Technical, scientific work, namely Engineering, Heavy industry, Construction, Health services and Public administration. Let us for these industries calculate the quotient between actual and estimated values, Q^t_{ijk}/P^t_{ijk}, which can be seen as a region-specific interaction effect between occupation group and industry. By analogy with other effects, this interaction effect is denoted λ_{ijk}.

It is interesting to note that in Stockholm this effect is positive (greater than one) for all industries except Public administration. The fact that the actual employment of Technical, scientific work in Public administration is only 75 percent of the estimated number could be a result of the abovementioned location policy. As to the private sectors the results give support to the assumption of a locational competitive advantage for R&D-activities in the metropolitan areas. In the framework of the product cycles theory, our interest is focalized on the manufacturing industries, here illustrated by Engineering in fig. 6.

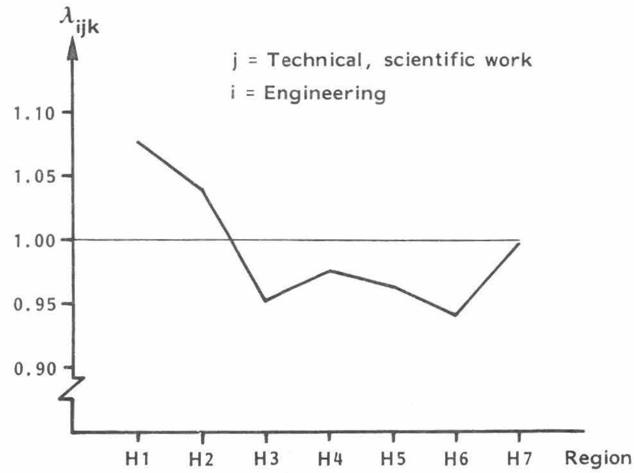


Fig. 6. Interaction effects between Technical, scientific work and Engineering 1970–1980 in seven H-regions.

Comparing the interaction effect between region and the occupation group, λ_{jk}, from fig. 5 (the continuous line) with the interaction effect between Engineering and Technical, scientific work given in fig. 6, and summing up, leads to the following tentative conclusions: whereas the regional distribution of R&D-activities in general has tended to equalize, R&D-activities in manufacturing tend to concentrate to metropolitan areas and to Stockholm in particular. We have suggested that the former tendency partly is an effect of public location policy, aiming at decentralizing the location for this kind of activities. The general location pattern is obviously related to the level of intraregional accessibility.

4. Occupation and industry structure in the Stockholm-region

In this section the entropy model will be used to analyse intraregional employment location. We will use both a geographical subdivision into a Core, Ring and Periphery and a division based on municipalities. By merging some municipalities and subdividing the municipality of Stockholm into five zones this gives 27 subregions.

Table 2 should serve as an introduction. It gives total employment in the Core, Ring and Periphery 1960–1980.

The table shows that employment decreased somewhat in the Core and increased rapidly in the Ring and Periphery. As shown in table 3 this decentralization of jobs reflects a decreased competitiveness of the Core.

Table 2
Employment changes in the Stockholm region 1960–1980 (1000's).

				Change in percent	
	1960	1970	1980	1960–1970	1970–1980
Core	376	360	357	−4	±0
Ring	140	196	236	40	20
Periphery	80	127	179	56	39
Total	593	676	760	14	12

Table 3
Employment structure 1960 and growth 1960–1980 attributed to competitive effects ($\exp \lambda_k$).

	1960	1960–1970	1970–1980
Core	1.92	0.83	0.89
Ring	0.72	1.25	1.04
Periphery	0.36	1.39	1.25

The first column gives the competitive effects initially and demonstrates the dominance of the Core. The second and third columns show that the Core lost and the Ring and Periphery gained in competitiveness during both the sixties and the seventies. The initial competitive effects should reflect historically created differences in access to input and product markets while the effects for the sixties and seventies should reflect changes in infrastructure and transportation technology that has brought about changes in the pattern of accessibility.

This reasoning suggests that further insight to employment location can be gained by comparing the effects of the entropy model with various measures of accessibility. A simple approach of this kind is illustrated in fig. 7, where the competitive effects for each of the 27 subregions in 1980 (estimated by means of eq. 10) have been plotted against car travel time to CBD in 1975. In cases of equal travel time the average competitive effect has been used.

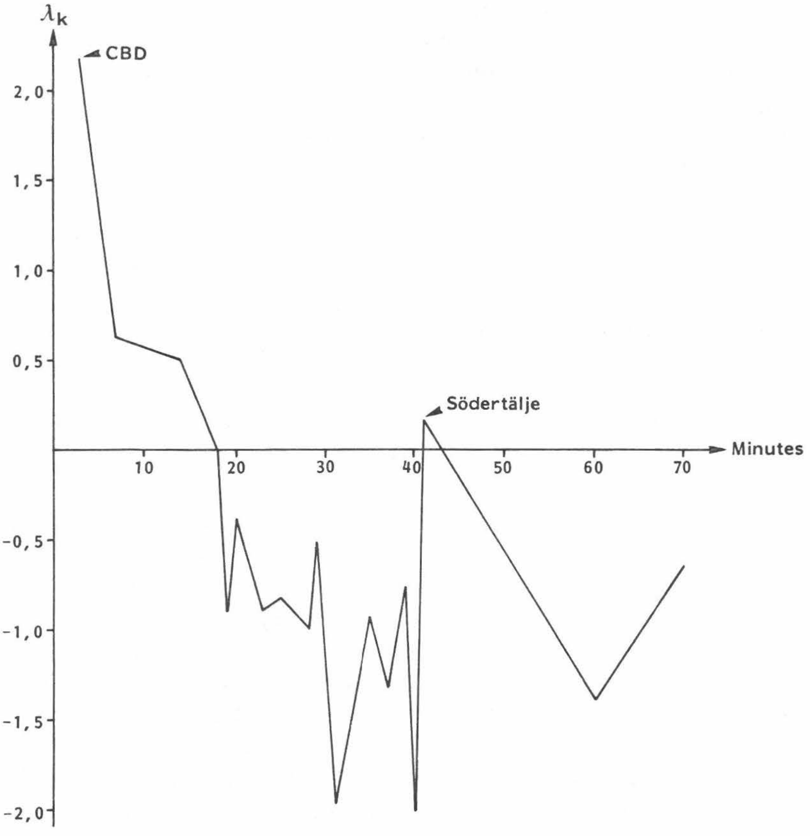


Fig. 7. Employment attributed to competitive effects in 1980 plotted against travel time to CBD by car in 1975.

Not surprisingly, the figure reveals a clear negative relationship between competitive effect and distance to the CBD. Except for the peak representing Södertälje, which is an old industrial center south of Stockholm, there is resemblance with a theoretically derived land price gradient, and it would be an interesting task to find out if the general form and various ups and downs of this curve [or rather a curve obtained by reformulating (10) in terms of densities] correspond to the actual land price gradient. In line with the Alonso bid price theory we should expect the steepness of bid-price curves to vary between different economic activities, see section 5. Such variations should be reflected in our interaction effects.

Considering the interaction between industries and subregions the results show that in 1960 Heavy industry, Engineering and Construction are smaller than expected in the Core and larger than expected in the Periphery. Light industry and National and regional service show a reverse pattern. The interaction effect for Engineering is larger in the Ring than in the Periphery but for all other industries the Ring falls in between the Core and the Periphery. Over time these intraregional differences have been maintained for Light industry and Construction, reinforced for Engineering and weakened for Heavy industry and National and regional services.

The trend for Heavy industry is somewhat surprising and may be due to head-office location to Stockholm from other parts of Sweden. It also reflects a rapidly increasing knowledge intensity. In general, the interaction effects between industries and Technical and scientific work look the same as at the national level. This means a higher than expected knowledge intensity in Heavy industry, Engineering, Construction, Public administration and Health service in 1960. Except for Public administration this knowledge orientation was strengthened during the sixties and seventies. Light industry, though not as knowledge-oriented as Heavy industry and Engineering, evidently exhibits a persisting centralizing pattern. Since it is dominated by printing and publishing a conjecture is that marketing and contacts with centrally located customers explain the central orientation.

In a study of intrametropolitan industrial location, Struyk and James (1975) found that manufacturing industries were clustered spatially to a significant degree and that each industry exhibited a distinctive locational pattern. Though a very crude subdivision of manufacturing is used here, our results point in the same direction. It is also evident that except for Agriculture and forestry this kind of pattern is less pronounced for other industries. This is illustrated in fig. 8 which shows the spatial interaction effects for Engineering and National and regional services as a function of travel time to CBD by car.

National and regional services follows the same downward sloping kind of pattern as the competitive effects whereas Engineering is characterized by lack of correlation with distance to CBD. Several factors can be suggested to explain the distinctive peaks for Engineering. The positive values in the outer

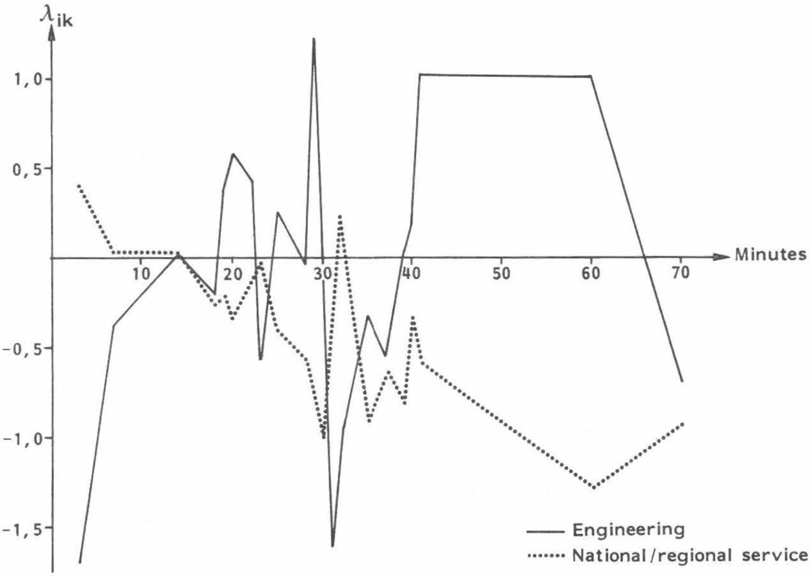


Fig. 8. Employment effects attributed to interaction between industry and subregion in 1980 as a function of travel time to CBD by car in 1975 (Engineering and National and regional services).

parts of the region are probably due to availability of docks and piers and to clustering with Heavy industry which exhibits positive effects in the same subregions. The peaks in the Ring may reflect that a number of areas in this part of the region are traditional sites of manufacturing activity. Since universities and highly educated labour are easily accessible from most subareas in the Ring, the high knowledge intensity of Engineering is another potential explanatory factor.

As for the interaction between occupations and subregions the results demonstrate a clear difference between knowledge-oriented and manufacturing occupations. In 1960 the Core dominates in knowledge-oriented and administrative work. Service and manufacturing work shows a reverse pattern. With the exception of Administrative work this picture seems to have been reinforced during the sixties and seventies. The trend is certainly weak and to some extent contradictory for Technical and scientific work and for Service work but it is clear-cut for Other knowledge-oriented work and for Manufacturing work.

The counteracting pattern for Technical and scientific work and Manufacturing work is further illustrated in fig. 9, which for each group gives the result of regressing the interaction effects on travel time to CBD by car.

Both relationships are clearly significant. Even though the determination coefficients are quite low, these results support the hypothesis that

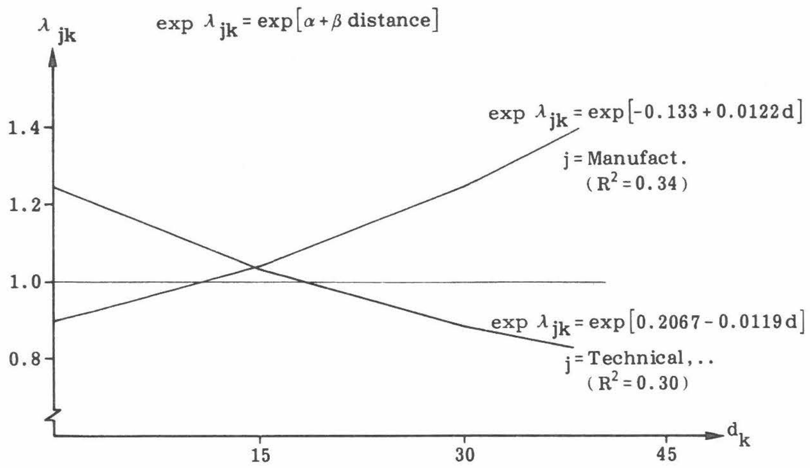


Fig. 9. Employment effects attributed to interaction between occupations and subregions in 1980 as a function of travel time to CBD by car in 1975 (Technical and scientific work and Manufacturing and agricultural work).

knowledge-intensive activities are more centrally located than manufacturing activities. These results are also in accordance with other analyses, e.g., Gera and Kuhn (1979).

As argued above, access to different kinds of labour should also be considered when analyzing the location pattern. In order to elaborate this idea further, the occupational pattern of the labour force in the subregions have been added as an explanatory factor in the regressions. The occupational structure of the labour force has been defined as the number of workers living in a subregion and having knowledge-oriented occupations divided by the corresponding number in other occupational groups. In order to allow a time-lag, 1970 data are used. The coefficient obtained when adding this variable in the regression for Technical and scientific occupations is positive and highly significant and the determination coefficient increases from 30 to 42 percent (adj R^2). The coefficient for distance increases somewhat but is still significantly negative.

We have also tested the ratio between the interaction effects for Technical and scientific work and Manufacturing work as a dependent variable. This results in a determination coefficient equal to 52 percent. If the non-significant intercept is excluded the coefficient increases to 65 percent.

All in all the outcome of the regressions confirms the usefulness of classifying employment according to occupations when analyzing intraregional location, and demonstrates the importance of considering the spatial-occupational pattern of the labour force. The latter has only been studied superficially as yet, but some results may still be of interest.

Classifying the labour force according to sub-region and occupational group and applying the entropy model, the results show clearly that the population also has decentralized 1960–1980. The competitive effects also reveal that this decentralization, like the decentralization of employment, was more rapid in the sixties than in the seventies.

However, the locational pattern of the labour force seems more stable than that of the employment. The chi-square values for the various applications with the entropy model show the following.

For the labour force the obtained chi-square values are lower if the 1960 and 1970 locational patterns are used when estimating the 1970 and 1980 patterns, respectively, than if this information is not utilized, i.e., if eq. (10) is used. The chi-square values associated with the employment estimations show a reverse pattern, i.e., chi-square is lower when the a priori information is not utilized. Since the differences in chi-square values are significant, it may tentatively be concluded that residential location is more inert than employment location or at least that information on the historical location pattern is more essential as regards the labour force compared to employment.

5. Intraregional location of employment and firms

The results from the previous section indicate that intraregional location of different activities could be related to the level of intraregional accessibility. Further, we have touched upon the significance of a changing infrastructure to explain a changing location pattern. In this section we will follow up these indications by estimating some explicit location models.

According to the Alonso bid-price theory the choice of location can be seen as a bidding process for land. For a given profit level each firm has a maximum bid-price for each location. The steepness of the bid-price curve varies between firms according to the depencence on a central location. Close to the centre economic activities will cluster if they have a limited need of land per unit of output combined with a large need of transportation and communication facilities. Other activities with less dependence of good accessibility will tend to locate more peripherally.

As a first approach we have used cross-section data to look upon the empirical relationships between land prices (assessments) and distance to CBD for firms in various industries. Data were obtained by combining samples from two official registers.

We have estimated the relationship between land price and distance to CBD for five industries. The results are illustrated in fig. 10.

As expected, the land price is falling as distance to CBD increases and the order between industries seems, by and large, to be in accordance with the theory. However, for each industry we can notice that a great deal of the

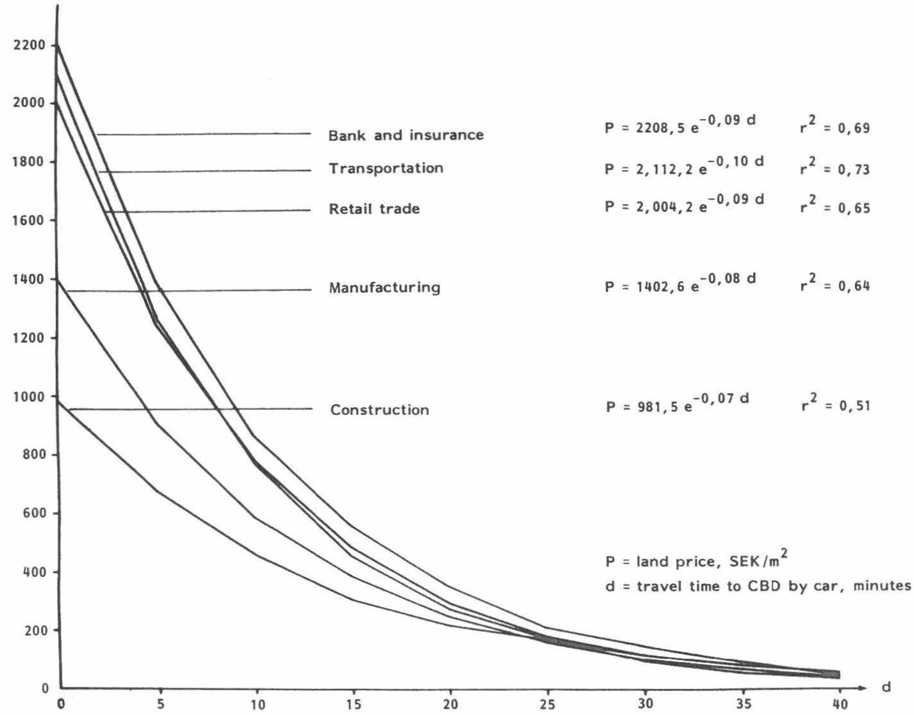


Fig. 10. Estimated relationships between land price and distance to CBD.

land price variation is not explained by distance to CBD, an observation in line with the pattern derived in the previous section. Further, the estimates for different industries do not produce a neat envelope but on the contrary a great land price variation at a given distance. It should be pointed out that the frequency of firms varies a lot between industries at each distance. It should also be stated that the municipalities play an important role in the land market, both as owners of land and being responsible for master planning.

Nevertheless, fig. 10 offers some points of interest for the following discussion. The order of the estimated land price gradients as between e.g., Bank and insurance and Manufacturing is in accordance with the theory; the former generally being more knowledge-oriented, more dependent upon communication and hence more dependent on good accessibility. Most firms are consequently located within or near CBD, whereas most firms in Manufacturing are located more peripherally. There are however atypical locations and not only because of locational disequilibrium (where firms have not relocated or otherwise adapted their activity to altered location conditions) but also because 'industry' is an inadequate definition of economic activities. By reasons put forward in previous sections we will in the following use both industrial and occupational definitions of economic activities.

The general model applied is a multinomial logit specification, based on the assumption of Weibull distributed random disturbance term, where the probability that a location unit of type t occupies a site k with characteristics X_k is

$$\Pr(k|t) = \exp[f_t(X_k)] \Big/ \sum_n \exp[f_t(X_n)]. \tag{11}$$

As shown by Anas (1981) the entropy and the multinomial logit (MNL) modeling approaches are identical. The MNL model, of the same type as Lee (1982), will be applied first to employment data and second to firms as location units.

In connection to the previous section we will first try to explain the location pattern of some selected activities, where the Stockholm region is subdivided into 40 zones (k) and where activities are defined by employment by industry and occupation group (t). The zonal attributes (X_k) to be used as independent variables are the following: First, there is one variable associated with accessibility to the labour market, ACC, defined as

$$ACC_k = \sum_n^{40} Q_{..n} \exp(-\alpha d_{kn}), \quad \text{where} \tag{12}$$

$Q_{..n}$ = employed dwelling population in zone n,
α = constant,
d_{kn} = travel time by car between zone and k and zone n.

It should be noted that ACC is highly correlated with CBD-distance. The rationale of including this variable has previously been hinted at. By differentiating between knowledge-oriented occupations and manufacturing occupations we should expect that the former are more dependent on a good accessibility.

Second, we have used two dummy-variables associated to local infrastructure; $MWAY$ representing motorway connections and $RAIL$ representing the presence of railway/subway stations in the zone. Third, there are two dummy-variables for zone; if the zone belongs to the $CORE$ or to the RING, previously defined. These zone dummies are introduced to catch various kinds of 'hidden' infrastructure capital, supposedly slowly changing, influencing employment location.

The model has been applied to employment in 1980 in three manufacturing industries: Engineering, Chemical industry and Other manufacturing (mainly Printing and publishing). For each industry employment has been twofold classified: one group of knowledge-oriented occupations (Technical, scientific and Other knowledge based work according to previous definitions) and the other group containing Manufacturing occupations. Additionally for Engineering we have used the corresponding employment figures in 1960, both as dependent and independent variables, to illustrate the changing location pattern. The estimation results are reported in table 4.

Considering that the models are simple and including only a few variables, the results in table 4 indicate an overall significance that is pretty good. With some exceptions the estimated coefficients have expected signs. For engineering, location of knowledge-oriented occupations is much more dependent on good accessibility than location of manufacturing occupations. The abovementioned decentralization since the sixties is borne out when comparing the coefficients of the zone-dummies in 1960 and 1980. The sign of *RING* is still positive in 1980 for knowledge-oriented occupations, which could be another indication of the importance of accessibility.

As for Chemical industry and Other manufacturing, interpreting the results is more complicated. Seemingly contrary to expectations, *ACC* has negative sign for knowledge-oriented occupations. On the other hand the coefficients for the *CORE*-dummy are strongly positive. This indicates a mixed location pattern, of both strong centralization and decentralized location, which partly could be explained by the mix of industries, e.g., refineries (in the periphery) and pharmaceutical industries (both central and peripheral sites). Other manufacturing is by its very name a heterogeneous mix of industries, where Printing and publishing is one sector with high dependence on good accessibility to customers located in the core. Judging from both *ACC* and the zone-dummies, a higher dependence on good accessibility for knowledge-oriented activities is a hypothesis which is given support. Further, the variables associated to local infrastructure have generally the expected positive influence upon employment location, although there is no reason for discrimination between occupations.

However, even as regards employment location, the model is certainly not complete; local infrastructure is roughly represented, no airports or harbours included for one thing. Further, since no distinction is made between employment in firms of different size, interpreting the results strictly is an awkward task. The estimated coefficients are probably seriously biased towards reflecting the location behaviour of firms with many employed.

Therefore, we will end this section by applying the MNL model to proper location units, i.e., firms. A survey has recently been carried out providing the data necessary for such an approach. The survey refers to 1981 and comprises in total 315 firms within manufacturing, reporting on a lot of firm-specific attributes. Combining this information with information of zone attributes as above we have applied the MNL model to two categories of

Table 4

Estimated coefficients[a] of logit model of employment location in three manufacturing industries in Stockholm region, 40 zones.

	Knowledge-oriented occupations					Manufacturing occupations				
	Engineering			Chemical	Other	Engineering			Chemical	Other
Variable	1960	1980	1980	1980	1980	1960	1980	1980	1980	1980
ACC	—	1.74	2.28	−1.07	−0.73	—	0.36	0.81	0.05	0.18
		(81.9)	(62.8)	(−14.2)	(−12.9)		(17.6)	(26.9)	(0.60)	(5.04)
MWAY	—	0.69	0.40	1.52	0.64	—	0.78	0.78	0.47	0.15
		(50.7)	(28.0)	(40.2)	(34.4)		(57.8)	(54.3)	(12.3)	(9.9)
RAIL	—	0.15	−0.14	1.58	1.24	—	0.83	0.79	0.67	0.36
		(8.07)	(−7.15)	(28.5)	(38.9)		(43.0)	(39.6)	(14.1)	(19.5)
CORE	3.64	—	−0.53	1.58	3.71	3.06	—	−0.16	0.44	1.69
	(1.66)		(−15.9)	(20.5)	(63.6)	(2.18)		(−5.2)	(5.34)	(48.6)
RING	2.19	—	0.51	0.18	1.48	1.48	—	−0.02	0.46	0.64
	(0.95)		(23.9)	(3.58)	(37.7)	(0.95)		(−0.91)	(8.85)	(27.1)
EMPL 1960	—	−0.28	—	—	—	—	0.80	—	—	—
		(−8.05)					(33.6)			
$2 \log L(\beta) - \log L(0)$	4.5	21 946	25 548	5298	36 926	5.9	17 466	16 454	1334	16 646
$\chi^2_{crit} = 0.05$[b]	5.99	9.49	11.1	11.1	11.1	5.99	9.49	11.1	11.1	11.1
ρ^2	0.68	0.46	0.54	0.43	0.64	0.66	0.42	0.39	0.19	0.42

[a] t statistics in parentheses.
[b] 5% chi-square value for 2 ($\log L(\beta) - \log L(0)$).

firms that are of a particular interest as regards the issue of knowledge intensity.

Engineering industry (113 firms) has been chosen being the major manufacturing industry in Stockholm and generally having a high knowledge intensity. All newly established firms (74 firms at present site after 1976) is the other group of firms. To make the choice-sets feasible we have reduced the number of alternative sites to the core, ring and periphery respectively. This means that the site attributes of the non-chosen alternatives have been calculated as the averages of the zones concerned in the respective alternative.

The zonal attributes to be used are the following: As an alternative to *ACC* we have included a variable reflecting the accessibility to employed dwelling population in knowledge-oriented occupations, *KNOWACC*, defined by the corresponding substitution for $Q_{..n}$ in (12). The reason for including this variable was hinted at in the previous section. The two zone-dummies are also included, as alternative specific constants. Finally, two firm specific variables (dummy-variables) are included: *WCOLL* representing a high share of white collar employment, i.e., non-manufacturing occupations, and *SPACE*, representing a high value of floor space per value added.

These two variables have been included since we have reasons to assume white collar oriented firms being more dependent upon good accessibility than firms that mainly are production units. As to *SPACE*, the reason is related: we expect that firms using a lot of floor space per unit output are out-bidded in central locations and hence should be located in the periphery. Both dummy-variables take the value one for two alternatives, core and ring, and the value zero for the alternative periphery.

Table 5 reports the results of estimations of the MNL model (11) where the accessibility variables are alternatively included.

Judging from the coefficient of *KNOWACC*, as compared to *ACC*, the accessibility to knowledge-oriented labour force is a variable of considerable importance for the choice probabilities. For newly established firms the χ^2 test shows that the estimated model (1) is not significantly different from the corresponding null model. In this case the small and non-significant coefficient for *ACC* indicates that general accessibility is a variable of small importance for new and relocated firms.

From the estimated models (2) we find that all zone-dummies have negative sign; thus indicating a general lower probability of central locations in the core and ring. However, as expected, white collar oriented firms, e.g., head offices, show a higher probability of central locations. As regards the *SPACE*-dummy, it does not show to be of any significance.

To conclude, the results indicate that accessibility to knowledge-oriented labour force is a variable of considerable importance when modeling the location behaviour of both kinds of firms. In the case of engineering we know

Table 5
Estimated coefficients[a] of logit model of firm location in 1981 for Engineering and newly established firms.

	Engineering firms		Newly established firms	
Variable	(1)	(2)	(1)	(2)
CORE	−1.17	−3.49	−0.34	−3.15
	(−1.30)	(−3.33)	(−0.30)	(−2.44)
RING	0.26	−0.91	0.06	−1.31
	(0.47)	(−1.51)	(0.09)	(−1.82)
ACC	1.51	—	0.41	—
	(1.38)		(0.30)	
KNOWACC	—	5.46	—	4.63
		(3.61)		(2.57)
WCOLL	2.54	2.69	1.57	1.77
	(2.42)	(2.55)	(1.96)	(2.17)
SPACE	−0.51	−0.43	0.15	0.44
	(−1.06)	(−0.88)	(0.25)	(0.72)
Number of firms	113	113	74	74
2 (log $L(\beta)-\log L(0)$)	37	49	10	17
$\chi^2_{crit}=0.05$[b]	11	11	11	11
ρ^2	0.15	0.20	0.06	0.10

[a]*t* statistics in parentheses.
[b]5% chi-square value for 2 (log $L(\beta)-\log L(0)$).

that this industry has a relatively high knowledge-intensity, and the results thus confirm what was hinted at in the previous section.

6. Conclusions

Stagnation and inner city problems have for some time been common attributes to characterize the metropolitan development in the industrialized countries. To some extent this gloomy picture has become contradicted by arguments stressing the importance of the metropolitan areas playing a seedbed role in the process of innovation and creativity.

In this paper we have tried to explore this seedbed role by decomposing the interregional and intraregional employment change by applying an entropy model. In addition to effects attributed to regions (zones), industries and occupations, interaction effects between region (zone) and industry/occupation have been derived. The results indicate that the metropolitan areas in general, and the Stockholm region in particular, have locational competitive advantages in knowledge-oriented activities. The intraregional analysis performed is also confirming the value of this kind of decompositions.

As for intraregional location, applications based on the Alonso type of theory demonstrate the importance of introducing different measures of accessibility. The location of labour force by different occupational groups seems to be of considerable importance for the location of knowledge-oriented activities.

References

Alonso, W., 1964, Location and land use (Harvard University Press, Cambridge, MA).
Anas, A., 1981, Discrete choice theory, information theory and the multinomial logit and gravity models, Paper presented at the North American Meetings of the Regional Science Association, Nov. 13–15, Montreal.
Andersson, Å. and B. Johansson, 1984a, Industrial dynamics, product cycles and employment structure, Working paper, WP-84-09 (IIASA, Laxenburg).
Andersson, Å. and B. Johansson, 1984b, Knowledge intensity and product cycles in metropolitan regions, Working paper, WP-84-13 (IIASA, Laxenburg).
Anderstig, C., 1981, Regional structural change and international specialization, Umeå Economic Studies 96.
Burns, L.S. and R.G. Healy, 1978, The metropolitan hierarchy of occupations, Regional Science and Urban Economics 8, no. 4.
Gera, S. and P. Kuhn, 1979, Occupation and job location patterns: An analysis of the Toronto census metropolitan area, Canadian Geographer 23.
Kuhn, S., 1983, Computer manufacturing in New England, Report from Joint Center for Urban Studies of MIT and Harvard University (Cambridge, MA).
Lee, K., 1982, A model of intraurban employment location: An application to Bogota, Colombia, Journal of Urban Economics 12.
Norton, R.D. and J. Rees, 1979, The product cycle and the spatial decentralization of American manufacturing, Regional Studies 13.
Paraskevopoulos, C., 1974, Patterns of regional economic growth, Regional Science and Urban Economics 4, no. 1.
Perloff, H. and L. Wingo, 1961, Natural resource endowment and regional economic growth, in: J.J. Spengler, ed., Natural resources and economic growth (Resources for the future, Washington, DC).
Premus, R., 1982, Location of high technology firms and regional economic development, Staff study prepared for the use of the subcommittee on Monetary and Fiscal Policy, U.S. (Washington, DC).
Snickars, F. and J. Weibull, 1977, A minimum information principle: Theory and practice, Regional Science and Urban Economics 7, nos. 1/2.
Struyk, R.J. and F.J. James, 1975, Intrametropolitan industrial location (Heath, Lexington, MA).
Theil, H. and R. Gosh, 1980, A comparison of shift-share and the RAS adjustment, Regional Science and Urban Economics 10, no. 2.

Regional Science and Urban Economics 16 (1986) 123–143. North-Holland

MODELLING THE LONG-TERM EFFECTS OF TRANSPORT AND LAND USE POLICIES ON INDUSTRIAL LOCATIONAL BEHAVIOUR

A Discrete Choice Model System*

Yoshitsugu HAYASHI, Tomohiko ISOBE and Yasuo TOMITA**

Nagoya University, Nagoya 464, Japan

Received May 1985, final version received August 1985

This paper describes a disaggregate behavioural model system developed for forecasting industrial locations. It is structured basically in terms of a nested logit model, covering relocation decisions, area-wide locational choices and local locational choices together with shipment destination choices. Several techniques are developed to overcome the difficulties in the application of discrete choice models to spatial problems. The model system was calibrated for the Nagoya metropolitan area in Japan and its validity was tested using another data set. It allows analysis of the effects of transport and land use policies not only by zone but also by firms of different attributes such as sector and capital size.

1. Introduction

Transport policy not only directly affects transport conditions and subsequent travel behaviour but, along with land use policies, also has significant implications for locational conditions and the subsequent location of economic activities. The latter effects take a much longer time to appear but are a fundamental determinant of the spatial structure of the region and consequently its economy and environment. In other words, once a wrong policy is implemented, it could take a long time to recover and to improve the regional economy and environment. The effect on industrial location is one of the essentials to be examined.

*This paper was presented at the International Conference on Transport Behaviour, 16–19 April 1985, Noordwijk, The Netherlands. An early version of this study was presented at the 8th Pacific Regional Science Conference, August 1983, Tokyo.

**The authors are indebted to Professor Shogo Kawakami and Yasuhiro Hirobata, Nagoya University, who gave us valuable suggestions in discussions. We are also grateful to John Roy, CSIRO, Peter Bonsall, University of Leeds and to anonymous referees for their valuable comments on the context of the paper. The early development of the model was supported by the research grant awarded by the Japanese Ministry of Science, Culture and Education. The rest of the study was done during the author's (Hayashi) stay at the University of Leeds, being supported by the research grant awarded by the Kajima Foundation.

Methodologies to assess the short-term effects (e.g., modal choice change) of transport policies by using disaggregate behavioural models are already numerous and well developed, and their usefulness has been appreciated by practitioners, presumably because, besides their technical advantage, they provide information about the effectiveness of policies on individuals with different attributes.

However, tools for assessing the long-term effects are much less developed and are still in their infancy. This seems mainly because of the following difficulties in applying such methods to a spatial choice problem [Lerman (1983)]:

(1) too large a number of choice sets of spatial alternatives,
(2) specification of the reasonable choice sets which are recognized by decision-makers among the full set,
(3) interaction between spatial choice and non-spatial choices (e.g., travel mode, dwelling type, etc.).

There are nevertheless some examples of disaggregate behavioural models for locational analysis. Most of them are residential location models [e.g., Anas (1983), Ben-Akiva and Palma (1983)], but industrial location is rarely treated [e.g., Lee (1982)]. However, considering the wide ranging locational preferences between individual firms in various sectors, the traditional aggregate approach is likely to be misleading in describing these locational patterns. In addition, because of technological innovation, change of preference within a firm is more rapid than that within a household to residential location. These problems are sufficient reason for attempting to develop disaggregate behavioural models for industrial location; indeed, quantitative models for industrial location are lacking even if aggregate models are taken into account [Boyce, Day and McDonald (1970), Nakamura, Hayashi and Miyamoto (1983)].

This study is an attempt to develop a disaggregate model system for forecasting industrial location which takes into account both a firm's attributes and the characteristics of alternative locations, including not only transport conditions but the other major factors in locational decision-making. This model system could be used for appraising locational behavioural changes of individual firms and changes of industrial location patterns within metropolitan areas as long-term effects of transport policy and related land use policies.

The study area is Aichi Prefecture which includes the dominant part of the Nagoya metropolitan area which is one of the leading industrial areas in Japan, and where plants belonging to various kinds of manufacturing firms are located. A questionnaire survey of firms which had recently relocated in the area was conducted to obtain data for the calibration of the sub-models on the basis of revealed preference.

This paper describes a brief summary of the questionnaire survey, model formulation, model estimation and the validity test.

2. Concept of the model system

2.1. Structure of the model system

Although industrial locational behaviour is simultaneous with the decision-making to select a specific site, it seems appropriate in modelling to disaggregate the decision into several stages; first decide whether to relocate or not, choose an area-wide zone, and then determine a site. According to this idea, the model system is composed of four sub-models, which formulate the following choice problems:

(1) relocation decision-making,
(2) location of suppliers and customers,
(3) area-wide location decision,
(4) local location decision.

The system simulates the locational behaviour of individual firms according to the flow as shown in fig. 1.

In the first step, locational demand, which consists of both new and relocation demands, is estimated. In the second step, the spatial probability density of shipment destinations and of places of purchase of materials and parts is determined. According to transport conditions for shipping and purchasing as well as the other locational conditions, the utility[1] of each locational choice set for individual firms is calculated, the probability of locational choice and consequently the number of firms deciding to locate one of their plants in each zone is then estimated.

2.2. Relocation sub-model

The relocation sub-model estimates the probability of relocation of existing plants and calculates the amount of relocation demand. When a firm decides to relocate one or more of its plants, it compares utility in the current location and expected utility in potential locations. If we assume it has nearly complete information, the expected utility in potential locations is obtained from the extreme value distribution of utilities in zones which are calculated by the locational choice sub-models. Thus, this sub-model is linked with the locational choice sub-models using a nested logit model.

[1]'Utility' is and will be used in this paper in its wider sense including the concept of profit.

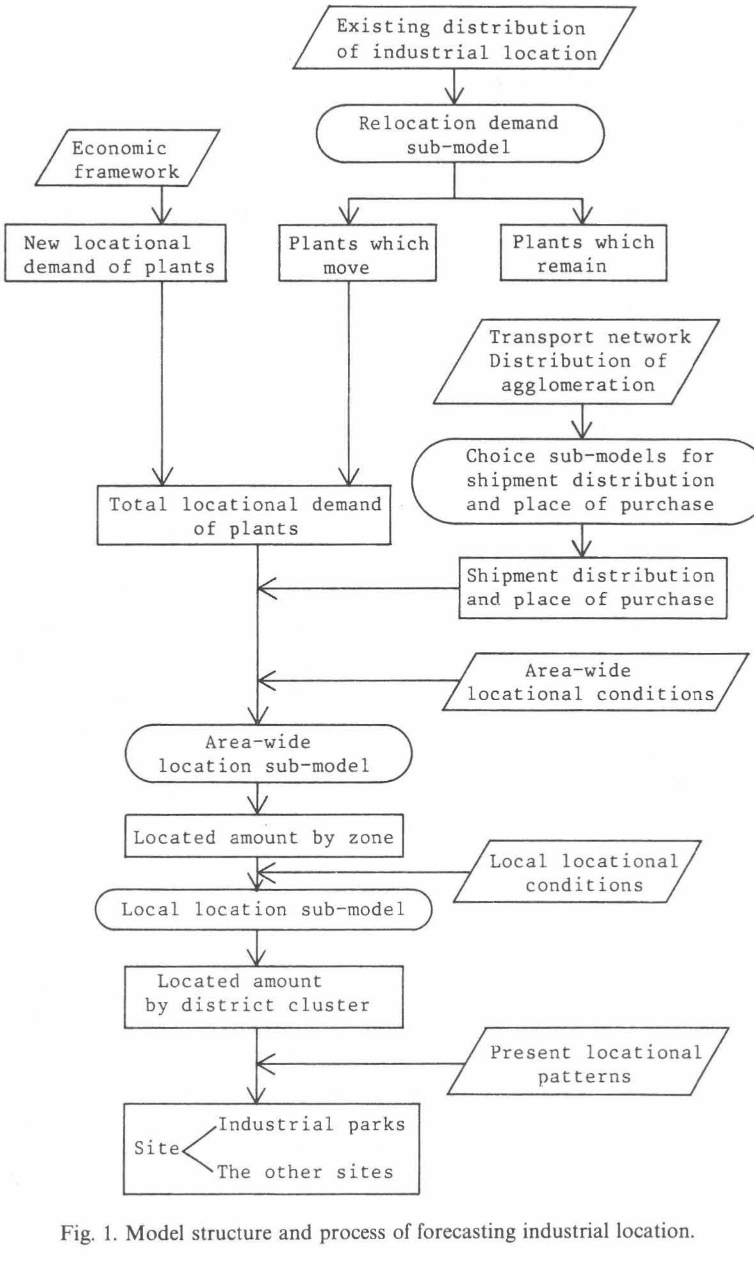


Fig. 1. Model structure and process of forecasting industrial location.

2.3. Locational choice sub-models

It is quite difficult to build a multinomial choice model which directly represents the choice between sites because of the large size of the choice sets. Therefore, the locational choice process is broken down into two stages which consist of area-wide choice and local choice, and these are connected by a nested multinomial logit model. This method guarantees the consistency between choices at the area-wide level and local level using composite variables. Such a disaggregation in terms of spatial level makes it easy to understand locational behaviours.

The hierarchical choice sets are nested as shown in fig. 2. The study area, Aichi Prefecture, is divided into 12 zones, each of which is further divided into 3 clusters of districts which consist of cities and rural municipalities as shown in fig. 3c.

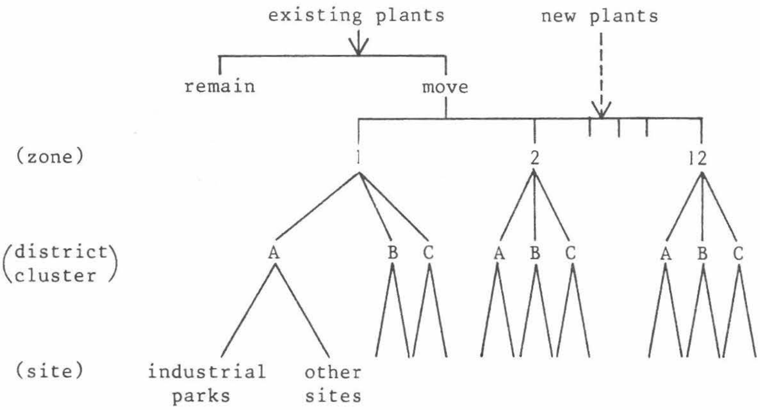


Fig. 2. Choice sets in relocation sub-model and locational choice sub-models.

At the district choice level it is very difficult to estimate a multinomial logit model regarding each district as a choice set, because it is a very large number. Therefore, to obtain a logit model, we classify districts into a small number of groups common to all zones using cluster analysis, according to local locational factors such as average land price, distance to the nearest expressway interchange and area of zoning for industrial use; here, principal component cluster analysis is used to consider influences of correlations between explanatory variables. The average characteristics of each categorized cluster of districts are shown in table 1.

The choice between industrial parks and the other sites within each district is taken directly from the survey data.

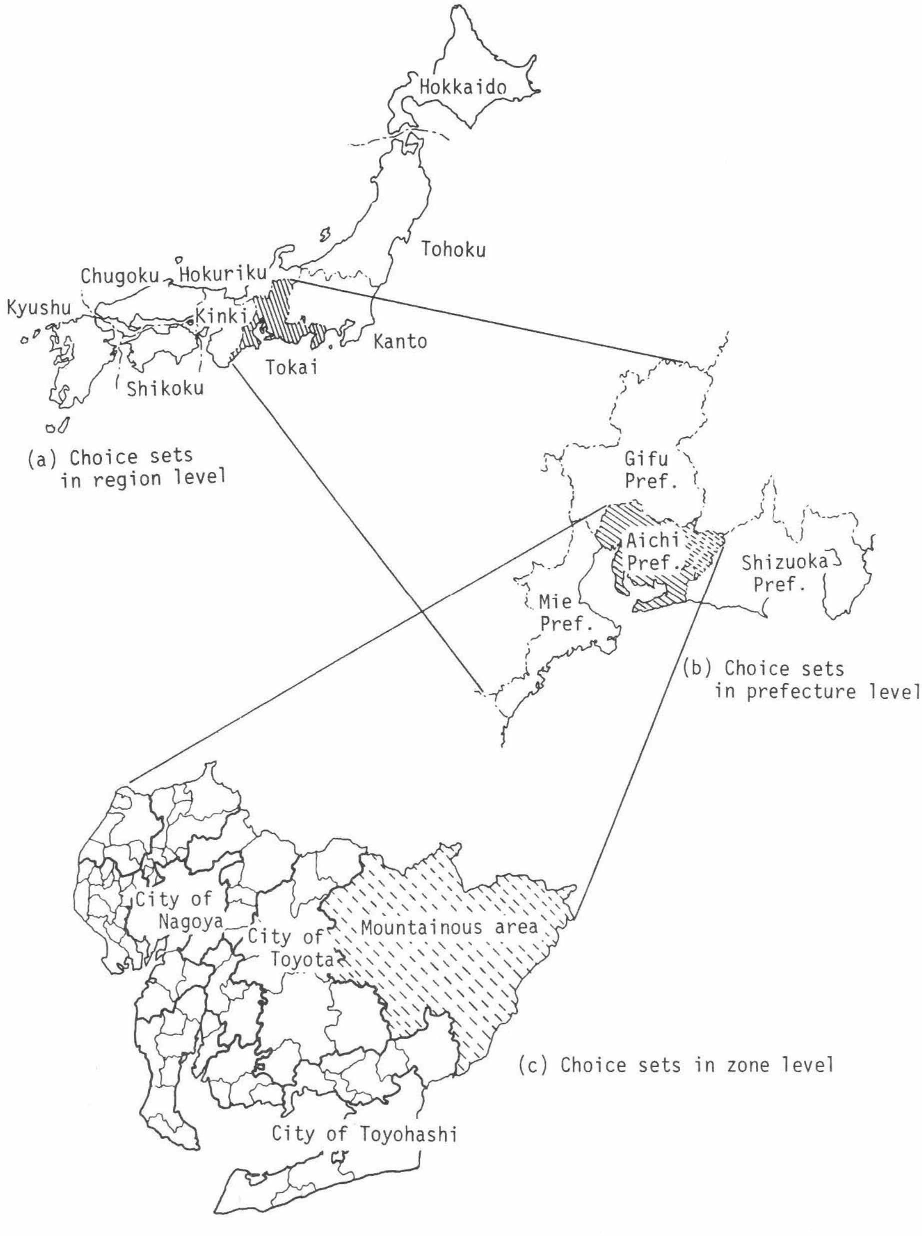


Fig. 3. Hierarchical choice set of space.

Table 1
Average characteristics of each district cluster.

District cluster	Area of zoning for industrial use in the cluster (km^2)	Distance to nearest interchange of expressway (km)	Land price (yen/m^2)
A	0.34	11.1	25,200
B	1.28	15.9	33,600
C	1.87	27.8	35,200

2.4. *Choice sub-models of shipment destinations and places of purchase of materials and parts*

In most existing empirical models for industrial location, firms' shipment destinations and places of purchase of materials and components are assumed to be exogenous. However, prediction errors in shipment destinations and places of purchase greatly influence the accuracy of the forecasts because transport time/cost for shipment and purchase is generally a dominant factor in locational decision-making. In residential relocation, each household has normally only one work place so that the data are comparatively easy to obtain, and it is not changed before and after relocation in most cases. However, in industrial relocation, firms generally have multiple shipment destinations and purchase places, so that the data are very hard to collect. In addition, these are sometimes changed before and after relocation. Therefore it is quite necessary to develop models for estimating locations of such places.

This model system equips the sub-models which choose shipment destinations and purchase places. The choice sets are nested hierarchically in space as shown in fig. 3, from nation-wide choice to local (district) choice.

3. Model formulation

Model formulations for each step of the locational decision-making are described in the following sections.

3.1. *Formulation of locational choice sub-models*

The utility of a location in zone i, district cluster j to a firm in industrial sector k is given as

$$U_{ij}^k = V_i^k + V_{ij}^k + \varepsilon_{ij}^k. \tag{1}$$

This equation states that the locational utility consists of three additively separable parts. The first part, V_i^k, measures the part of locational utility due to the area-wide location factors which vary between zones. These would be such characteristics as the transport cost for shipment, the transport cost for purchase, the commuting conditions of employees, availability of labour force, etc. The second part, V_{ij}^k, expresses local location factors such as distance to the nearest interchange of expressways, land price, amount of supply of industrial sites, etc. The third part of locational utility, ε_{ij}^k, is a random variable due to unknown (unobserved) characteristics, including preference differences of individual firms, random effects and errors in measurement. The probability that a firm k will choose an alternative place (i, j) to locate its plant is given as

$$P_{ij}^k = \text{Prob}\,[U_{ij}^k > U_{i'j'}^k, (i', j') \neq (i, j)]. \tag{2}$$

The specific form of (2) depends on what is assumed about random error terms, ε_{ij}^k. If the disturbance, ε_{ij}^k, is assumed as independent and identically Gumbel distributed, P_{ij}^k can be computed as the nested multinomial logit model. Because locational utility is additively separable, the probability can be described as

$$P_{ij}^k = P_i^k \cdot P_{j/i}^k. \tag{3}$$

Here $P_{j/i}^k$ is the conditional probability that the firm will choose district cluster j to locate its plant given that zone i has been chosen and P_i^k is the marginal probability that zone i will be chosen. These probabilities are described by the following formulations:

$$P_{j/i}^k = f(S_j)\exp(V_{ij}^k) \Big/ \sum_{j'=1}^{J} f(S_{j'})\exp(V_{ij'}^k),$$

$$i = 1, \ldots, I; \quad j, j' = 1, \ldots, J, \tag{4}$$

$$P_i^k = \exp[V_i^k + \alpha_0 v_i] \Big/ \sum_{i'=1}^{I} \exp[V_{i'}^k + \alpha_0 v_{i'}], \tag{5}$$

$$v_i = \log \sum_{j'=1}^{J} \exp(V_{ij'}), \tag{6}$$

where S_j denotes the area zoned for industrial use within district cluster j. v_i is the composite value given as the expected value of the extreme distribution of locational utility for all district clusters within zone i. α_0 is the measure of the similarity of district clusters in the same zone with respect to their unobserved attributes.

Eq. (4) states that the probability of choosing district cluster j given the choice of zone i is calculated by a multinomial logit model, and thus depends on the relative locational utility of the district clusters keeping zonal characteristics constant. $f(S_j)$ is a generalized correction factor of choice probability which represents not only the size effect of choice set [Anas (1983)] but also the combined effects of capacity constraint of district cluster j for industrial use and attractiveness due to the size of industrial agglomeration. The form of $f(S_j)$ is determined due to the mutual intensity of each effect. Eq. (5) is the marginal zone choice probability for a nested logit model. The probability of choosing a zone is a function of the locational utility of the zone plus the composite value of locational utilities of the district clusters within the zone. Eq. (6) gives the composite value of locational utilities in the local location sub-model. ε_{ij}^k is the measure of district clusters in the same zone with respect to their unobserved characteristics.

3.2. Formulation of the relocation sub-model

The utility of relocation to a firm in industrial sector k of its plant being located in zone i_0 is given as

$$U_{\mathrm{m}}^{ki_0} = v_{\mathrm{m}}^{ki_0} + \varepsilon_{\mathrm{m}}^{ki_0}. \tag{7}$$

This equation states that utility consists of two additively separable parts. The first part, $v_{\mathrm{m}}^{ki_0}$, measures the part of utility due to relocation, which is a composite value given as the expected value of the extreme distribution of locational utilities for all available zones. That is

$$v_{\mathrm{m}}^{ki_0} = \log \sum_{(i,j)\in R_{i_0}} \exp(V_{\mathrm{m}i}^k + V_{\mathrm{m}ij}^k) \tag{8}$$

where R_{i_0} denotes a set of relocatable zones available from zone i_0 and $V_{\mathrm{m}i}^k + V_{\mathrm{m}ij}^k$ is previously defined in eq. (1). The second part, $\varepsilon_{\mathrm{m}}^{ki_0}$ is a random variable due to unknown (unobserved) characteristics, including preference differences of individual firms, random effects and errors in measurement. The probability that a plant belonging to firm k being located in zone i_0 will be relocated (m) is given as

$$P_{\mathrm{m}}^{ki_0} = \mathrm{Prob}\,[U_{\mathrm{m}}^{ki_0} > U_{\mathrm{s}}^{ki_0}], \tag{9}$$

where m denotes relocation while s indicates remaining (non-relocation). The specific form of (9) depends on what is assumed about the random terms, $\varepsilon_{\mathrm{m}}^{ki_0}$. If the disturbance, $\varepsilon_{\mathrm{m}}^{ki_0}$, is assumed as independent and identically Gumbel distributed, $P_{\mathrm{m}}^{ki_0}$ is computed as the nested logit model. Because relocation utility is additively separable, the probability can be described as

$$P^{kio}_{mij} = P^{kio}_{m} \cdot P^{kio}_{i/m} \cdot P^{kio}_{j/mi}. \tag{10}$$

Here $P^{kio}_{j/mi}$ is the conditional probability that the firm will choose district cluster j for the location of its plant, given zone i in the case of relocation (m). $P^{kio}_{i/m}$ is the conditional probability that the firm will choose zone i in the case of relocation (m) and P^{kio}_{m} is the marginal probability that relocation is chosen instead of remaining (s).

3.3. *Formulation of choice sub-models of shipment destinations and purchase places of materials and parts*

Shipment destination and purchase place of materials/parts are generally distributed nation-wide. Therefore, the choice structure for these destinations is assumed in terms of spatial hierarchy; region level, prefecture level and zone level (see fig. 3). As it is quite difficult to collect data representing locational utilities of zones in the other prefectures or regions outside the study area, a nested logit model cannot be applied here, but we can develop separate multinomial logit models for each of the three spatial levels of choice.

The sub-models are formulated based on the following ideas (see fig. 4): Each plant generally has multiple shipment destinations/purchase places (fig. 4a). Normally we can observe the aggregated volumes of freight (T^k_{il}) to multiple destinations l by individual plant of firm k. These data relate, of course, to firms' behaviour but not to discrete (0, 1) choices and cannot therefore be formulated by discrete choice models. However, by disaggregating a plant into its component activity units [all of which share the plant's attributes but each of which makes discrete (0, 1) destination choices as if it were an individual decision-maker (fig. 4b)], the volume T^k_{il} in interpreted as the aggregation of activity units of firm k which chooses destination l. Thus, we can formulate destination choice as a discrete model for activity units as shown below.

Given the zone i as location of plant of firm k, the utility of destination l for an activity unit (u:$u = 1, \ldots, N$) of the plant is given as

$$U^{ku}_{il} = V^{ku}_{il} + \varepsilon^{ku}_{il}. \tag{11}$$

This equation states that utility consists of two additively separable parts. The first part, V^{ku}_{il}, means the observable part of utility due to choice of destination l. The second part, ε^{ku}_{il}, is a random variable due to unknown (unobserved) characteristics.

The probability that an activity unit u of the plant would be attracted to destination l is given as

$$P_{il}^{ku} = \text{Prob}\ [U_{il}^{ku} > U_{il'}^{ku}, l \neq l']. \tag{12}$$

Assuming ε_{il}^{ku} as independent and identically Gumbel distributed, P_{il}^{ku} is described as

$$P_{il}^{ku} = \exp(V_{il}^{ku}) \Big/ \sum_{l'} \exp(V_{il'}^{ku}). \tag{13}$$

As the attributes of each activity unit are the same as those of the plant, the spatial distribution of destinations for the plant can also be calculated by the following equation:

$$P_{il}^{k} = \exp(V_{il}^{k}) \Big/ \sum_{l'} \exp(V_{il'}^{k}). \tag{14}$$

Provision of (0, 1) discrete choice data for estimating the above model [eq. (13)] is conducted according to the following procedures: Securing equal weights between plants $(k_1, \ldots, K)$ in the estimation, T_{il}^k is first normalized to be N_{il}^k subject to

$$\frac{N_{il}^k}{N_{il'}^k} = \frac{T_{il}^k}{T_{il'}^k}, \qquad l \neq l', \quad \text{and} \tag{15}$$

$$\sum_l N_{il}^k = N, \tag{16}$$

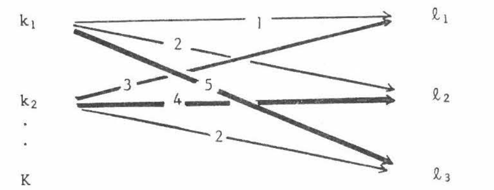
where N denotes number of imaginary activity units of each plant and can be any common multiple number of T_{il}^k $(l = 1, \ldots, L)$; [compare figs. 4c and 4d]. Disaggregating each plant into N activity units which are assumed as individual decision-makers, (0, 1) discrete shipment data can be obtained as shown in fig. 4e.

4. Empirical estimation

In this section we briefly describe the survey of plant location. The results for the specified logit models are then discussed.

4.1. Questionnaire survey of the determinants of plant location

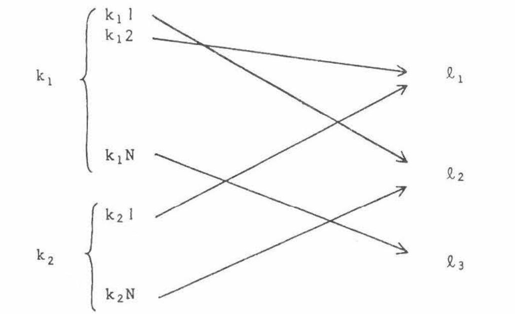
The survey was conducted for 416 plants which located for the first time in



*) $T^{k_1}_{i\ell_1} = 1,\quad T^{k_1}_{i\ell_2} = 2,\quad T^{k_1}_{i\ell_3} = 5,$

$T^{k_2}_{i\ell_1} = 3,\quad T^{k_2}_{i\ell_2} = 4,\quad T^{k_2}_{i\ell_3} = 2$

(a) Destination choices with different magnitudes



(b) Imaginary (0,1) discrete choices of destination by activity unit

Plant \ Destination	ℓ_1	ℓ_2	ℓ_3	Total
k_1	1	2	5	8
k_2	3	4	2	9
.				.
.				.
.				.
.				.

(c) Shipment volume ($T^k_{i\ell}$)

Plant \ Destination	ℓ_1	ℓ_2	ℓ_3	Total (N)
k_1	9	18	45	72
k_2	24	32	16	72
.				.
.				.
.				.
.				.

(d) Normalized shipment volume ($N^k_{i\ell}$)

Unit \ Destination	ℓ_1	ℓ_2	ℓ_3
$k_1 1$	0	1	0
$k_1 2$	1	0	0
.			
.			
$k_1 72$	0	0	1
Total	9	18	45

(e) (0,1) shipment choice

Fig. 4. Generation of (0, 1) discrete choice data from different magnitude of choice with an example.

Aichi Prefecture between 1976 and 1981. The number of samples with valid answers is 116.

The main items of the questionnaire were as follows:

(1) the date when the plant was established in its current location,
(2) the locations of both newly established and relocated plants (with the previous locations of movers),
(3) amount of electricity and water used,
(4) destinations and amount of shipment of major items of production of each plant,
(5) places and the quantity of purchase of major materials and parts of each plant,
(6) locational subsidies by local government which were considered by firm in selecting the site,
(7) reasons and main factors considered in location decisions,
(8) alternative sites and zones for the plant which the firm rejected, and its reasons,
(9) process of decision-making of location.

Tables 2 and 3 show the configurations of reasons for area-wide and local locational choices, respectively, as simple totals of questionnaire survey.

In area-wide locational decision-making (see table 2), 'accessibility to suppliers and customers' is one of the dominant factors, as has been suggested in

Table 2
Configuration of reasons for area-wide locational choice by industrial sector.[a]

	Reasons						
Sectors	A	B	C	D	E	F	G
Pulp and paper, ceramic	48.3	1.7	0.0	5.2	1.7	39.7	3.4
Food	50.0	3.1	0.0	6.3	6.3	28.1	6.2
Textile	44.7	6.6	0.0	9.2	6.6	31.6	1.3
Electrical machinery, precision machinery	37.5	8.3	0.0	8.3	0.0	33.3	12.6
Processing and general machinery	37.6	3.8	0.6	15.3	1.9	38.2	2.6
Plastics, rubber	40.5	2.7	0.0	8.1	2.7	45.9	0.1
Basic	46.9	4.2	4.2	14.4	2.1	28.1	0.1
Other manufacture	40.0	0.0	0.0	0.0	0.0	60.0	0.0
All sectors	48.1	5.3	1.1	8.4	2.3	32.8	1.9

[a]A = accessibility to suppliers and customers, B = availability of labour force, C = availability of water supply, D = subsidy by local government, E = special linkage with other firms, F = foundation place of firm or owner's personal connections, G = other reasons.

Table 3
Configuration of reasons for locally locational choice by industrial sector.[a]

Sectors	Reasons A	B	C	D	E	F	G
Pulp and paper, ceramic	17.2	8.6	37.9	24.1	3.4	3.4	5.4
Food	40.6	6.3	18.8	21.9	3.1	3.1	6.2
Textile	15.8	1.3	31.6	39.5	5.3	3.9	2.6
Electrical machinery, precision machinery	20.8	8.3	29.2	16.7	0.0	12.5	12.5
Processing and general machinery	17.8	8.3	34.4	29.9	1.9	6.4	1.3
Plastics, rubber	13.5	0.0	35.1	48.6	0.0	2.7	0.1
Basic	20.8	2.1	25.0	45.8	2.1	2.1	2.1
Other manufacture	40.0	0.0	60.0	0.0	0.0	0.0	0.0
All sectors	16.7	6.2	37.1	30.9	2.3	2.7	4.2

[a]A = accessibility to transport facilities, B = convenience for journey to work, C = land price or ease to purchase a site, D = advisory service by local government for suitable sites such as industrial parks, E = special linkage with other firms, F = owner's personal connections, G = other reasons.

traditional location theories. 'Foundation place of firm or owner's personal connection' is the other dominant; such kind of factors are not easy to represent by a quantitative model. 'Subsidy by local government' is a significant reason in 'Processing and general machinery industry' and 'Basic industry' which need large sites and large-scale capital equipment on which big amounts of property tax are imposed.

In local locational decision-making (see table 3), configurations are more dispersed than in the area-wide choice. But, 'land price or ease to purchase a site' and 'advisory service by local government' are dominant factors. The latter is quite a soft factor, which has not been treated in the traditional theory, and is more significant than 'accessibility to transport facilities' in more than half of the sectors. 'Food industry' is particularly sensitive to 'accessibility to transport facilities'.

According to these patterns of reasons for locational decision-making, manufacturing industries are classified into four types in this study as shown in fig. 5.

4.2. Estimation of locational choice sub-models

4.2.1. Local location sub-model

The local location sub-model estimates probability of choosing district cluster (A, B, C) within a zone (see table 1). A linear utility function is

- (a) Processing industries
  - General machinery
  - Electrical machinery
  - Transport machinery
  - Precision machinery
  - Manufacture of fabricated metal product
- (b) Urban consumer oriented industries
  - Miscellaneous manufacture
  - Furniture
  - Apparel
  - Leather
  - Rubber
  - Publishing and Printing
- (c) Local resource oriented industries
  - Textile
  - Lumber and Wood Products
  - Food
  - Ceramic
- (d) Basic industries
  - Pulp and Paper
  - Non-ferrous metal
  - Iron and Steel
  - Chemical
  - Petroleum and Coal

Fig. 5. Classification of industrial sectors.

assumed in this sub-model. Table 4 shows the explanatory variables and the estimated result of the model. The variables have been selected according to the results of the questionnaire survey mentioned in the previous section.

The first variable (X_1) represents local transport condition in terms of distance to nearest interchange of expressway. To distinguish the different marginal utilities between sectors, differential coefficients are assumed in preference to the dummy variables often adopted in the existing logit models. The estimated coefficients show that 'Processing industry' indicates the highest marginal disutility while 'Urban consumer oriented industry' the lowest.

The second variable (X_2) is area of zoning for industrial use. The coefficients are also differentiated by sector. They involve several meanings:

(1) requirement for industrial zoning for environmental reasons,
(2) orientation to industrial agglomeration,
(3) the effect of size of available areas in alternative district clusters [see eq. (4)].

The estimated coefficients show that 'Basic industry' indicates the highest marginal utility, which may be due to environmental considerations and for large industrial sites, while 'Urban consumer oriented industry' indicates the lowest presumably because of both its less needs for area due to its intensive land use and its orientation more to urban (consumer and tertiary industrial) agglomerations.

Table 4
Estimated results of local location sub-model.[a]

Variables descriptions	Categories	Variables	Coefficients	t-statistics
Distance to nearest interchange of expressway (km)	a	x_1^1	−0.073	4.1
	b	x_1^2	−0.0013	0.7
	c	x_1^3	−0.035	2.0
	d	x_1^4	−0.036	1.1
Area of zoning for industrial use (ha)	a	x_2^1	0.0027	3.4
	b	x_2^2	0.0017	1.4
	c	x_2^3	0.0025	3.4
	d	x_2^4	0.0034	3.0
Proportion of purchase cost of a site in firm's capital	With subsidies	x_{31}	−0.032	2.3
	Without subsidies	x_{32}	−0.085	1.9
Number of samples			116	
Number of choice sets			3	
Percentage of correctly estimated samples			76.7%	
ρ^2 statistic			0.359	

[a]a, b, c, d correspond to industrial sectors shown in fig. 5.

The third variable (X_3) represents ease of purchase of its site in terms of proportion of purchase cost in the firm's capital. Coefficients differentiate between the district clusters where locational subsidies by local government can and cannot be obtained. The estimated results show as expected that the subsidies make the coefficient less steep. The difference between the two coefficients implies the effect of subsidy.

The goodness-of-fit of calibration is comparatively high; the proportion of the correctly estimated samples is 76.7% and ρ^2-statistic is 0.359 (number of samples is 116).

4.2.2. *Area-wide location sub-model*

The area-wide location sub-model estimates the probabilities that a firm will choose a zone among 12 zones within a prefecture, and then the amount of location in each zone given the composition of firms' attributes. According to the results of analysis of the survey, major location factors are transport conditions to shipment destinations and from purchase places, availability of labour, locational subsidies by local government, etc., so that the indicators shown in table 5 are adopted as explanatory variables of the model. A linear utility function is assumed in this sub-model. Since the marginal disutilities of transport costs for shipment and purchase would differ between industrial

Table 5
Estimated results of area-wide location sub-model.[a]

Variables descriptions	Categories	Variables	Coefficients	t-statistics
Average transport time from zone of purchase of materials and parts (minutes)	a	X_4^1	−0.015	3.0
	b	X_4^2	−0.0091	1.0
	c	X_4^3	−0.017	3.3
	d	X_4^4	−0.020	2.0
Employee density of primary industrial sector in zone (persons per km^2)		X_5	0.0011	1.3
Value of composite variable representing total utility in local location sub-model		v	0.46	9.8
Number of samples				116
Number of choice sets				12
Percentage of correctly estimated samples				72.8%
ρ^2 statistic				0.583

[a] a, b, c, d correspond to industrial sectors shown in fig. 5.

sectors because of the difference of values per unit of freight, variables with differential coefficients are indexed by industrial sector.

X_4^1–X_4^4 represent transport time from zone of purchase of materials/parts by sector. According to the estimated result, the coefficients for transport cost show that 'Basic industry' indicates the most significant value followed by 'Local resource oriented industry' and 'Processing industry', and 'Urban consumer oriented industry' the lowest. X_5 represents the availability of local labour force in terms of employees density of primary industry. 'v' represents difference of local locational condition in terms of composite utility obtained from the local location sub-model and shows the highest contribution among the explanatory variables. Besides the above variables, an attempt was made to include 'transport time to shipment destination' and 'distance from the previous location of the plant', but since they had no statistical significance, they were excluded.

The goodness of fit of calibration is shown as 72.8% in terms of the proportion of the correctly estimated and 0.583 in terms of ρ^2 statistic (number of samples is 116). This result can be regarded as fairly good, considering the large number of choice sets.

By these sub-models locational choice problems of firms are described quantitatively and hierarchically in space, considering the major locational factors in traditional industrial location theory, such as transport conditions, labour force conditions and agglomeration conditions.

Table 6
Estimated results of choice sub-model for purchase places of material/parts at zone level.

Variables description	Variables	Coefficients	t-statistics
Average transport time to purchase places of materials/parts (minutes)	$\log X_6^1$	−0.64	11.0
Number of firms in zone of purchase of materials and parts	X_7^1	0.00016	23.3
Number of samples		62	
Number of choice sets		12	
Correlation coefficient		0.858	
ρ^2 statistic		0.308	

4.3. Estimation of choice sub-models of shipment destinations and purchase places of materials/parts

The sub-models estimate probabilities of spatial choice of shipment destinations and purchase places, which will be used as the weight for calculation of average transport time for shipment and purchase in the utility function of the area-wide location model. Here a choice model of purchase places estimated for 'Processing industry' at the zone level is shown as an example. A linear utility function is assumed also in this sub-model. As explanatory variables, the indicators shown in table 6 are adopted, where X_6 represents the average transport time to each given zone from all purchase place zones weighted by their amounts of shipment; X_7 represents industrial agglomeration.

The goodness-of-fit of calibration is comparatively high; 0.308 in terms of ρ^2 statistic and 0.858 in terms of correlation coefficient for spatial distribution of purchase place, given the zone of location, which would be a more adequate indicator of the goodness-of-fit in this case than the proportion of the correctly estimated activity units.

5. Validity test

The validity of the model is tested using a data set which is independent of those used for calibration. This was obtained from a governmental study which conducted a questionnaire survey of all the plants located in the Tokai region (see fig. 3a) while data used for calibration contains only plants with sites larger than $1000\,\mathrm{m}^2$. The data is classified by the year of location (before or after 1965) and by industrial sector.

Tables 7 and 8 show the proportion of samples whose district cluster and zone were correctly forecast by the local location sub-model and the area-wide location sub-model respectively.

Table 7
Results of validity test of local location sub-model.[a]

	Industrial sector				
Located year	Processing	Urban consumer oriented	Local resource oriented	Basic	All sectors
1926–1965	0.31 (10/32)	0.67 (205/307)	0.28 (13/46)	0.40 (4/10)	0.59 (232/359)
1966–1981	0.51 (29/56)	0.48 (46/96)	0.70 (124/177)	0.33 (19/58)	0.56 (218/387)
1926–1981	0.44 (39/88)	0.62 (251/403)	0.61 (137/223)	0.34 (23/68)	0.58 (450/782)

[a]The numbers show percentages of correctly estimated samples. Fractions in parentheses show ratios of the correctly estimated among the sample.

Table 8
Results of validity test of area-wide location sub-model.[a]

	Industrial sector				
Located year	Processing	Urban consumer oriented	Local resource oriented	Basic	All sectors
1926–1965	0.59 (55/93)	1.00 (12/12)	0.69 (35/51)	0.29 (8/28)	0.60 (110/184)
1966–1981	0.83 (10/12)	1.00 (4/4)	0.22 (2/9)	0.71 (5/7)	0.65 (21/32)
1926–1981	0.62 (65/105)	1.00 (16/16)	0.62 (37/60)	0.37 (13/35)	0.61 (131/216)

[a]The numbers show percentages of correctly estimated samples. Fractions in parentheses show ratios of the correctly estimated among the sample.

The overall percentages of the correctly estimated are 58% in the local location sub-model and 61% in the area-wide sub-model. Examining by sector 'Urban consumer oriented industry' shows the best fit, especially in the area-wide location sub-model and 'Basic industry' the worst. Except for 'Basic industry', the results seem fairly good by the standard of industrial location forecast.

The plants located during 1926–1965 are represented fairly well by sub-models which are calibrated by the data including plants which located during 1976–1981. This may be because these plants seem to have located in the sites with good locational conditions even today, for they have not relocated but have continued to stay there.

The results also reflect the difference of the size ranges of plants between the two data sets as described above. If the ranges were matched, the better validity could be obtained.

6. Conclusions

This paper is only a re-expression of industrial locational behaviour from the viewpoint of traditional location theory, in terms of random utility model. It is not a final result of modelling of industrial locational behaviour but rather a primary trial. Some important factors and behavioural norms are still excluded from the model system. For example, as the results of the questionnaire survey suggest, a firm's behaviour might reflect institutional factors, personal connections of its owner, or convenience of its organization. It is also important to note that, because of data limitations, the study has not involved new high-technology industries which might be more footloose and have a more area-wide (international) choice set and regrettably were not estimated – again because of data limitation on the basis of revealed preference.

However, the authors believe the experience which has been made up to this stage of the study could involve several new findings and suggestions which may be useful for practical policy evaluation and further development of modelling. We conclude the paper with a summary of the main features of this study.

(1) As locational behaviours are explicitly expressed at the level of the individual firm in terms of random utility theory, the model can provide more detailed information than most existing aggregate models, since it allows the analysis of effects of policies not only on zones but also on individual firms.

(2) The whole model system is structured basically in terms of a nested logit model corresponding to the real process of locational decision-making so that the interpretation is easy.

(3) The model considers not only transport conditions but other locational factors such as labour availability, land price, government subsidy, etc. and the attributes of firms such as industrial sector and capital size of the firm so as to distinguish the effects caused by different policies.

(4) To summarize, the model can test the effects of the following policies:

(a) road infrastructure improvement schemes,
(b) traffic management policies,
(c) locational subsidization policies by local government for different sectors and for different capital sized firms,
(d) policies of zoning regulation for industrial use,
(e) new industrial park developments.

(5) The following techniques were proposed to overcome difficulties in the application of discrete choice models to spatial problems and they provide good fits:

- to aggregate the large number of spatial choice sets into a manageable size, cluster analysis was introduced;
- to overcome the problem of small sample numbers, which usually happens in collecting behavioural data on industrial plants, differential coefficients by sector were assumed for several locational factors while common ones to all sectors were assumed for the others;
- a technique to formulate different magnitudes of choice in terms of discrete choice models was developed.

(6) The validity of the model was tested using a data set which is independent of those used for calibration, and a comparatively good result was obtained.

References

Anas, A., 1983, Dynamic forecasting of travel demand, residential location and land development: Policy simulations with the Chicago area transportation/land use analysis system, Paper presented at the International Symposium on New Directions in Urban Modelling (University of Waterloo, Waterloo).

Ben-Akiva, M. and A. Palma, 1983, Modelling and analysis of dynamic residential location choice (McMaster University, Hamilton).

Boyce, D.E., N.D. Day and C. McDonald, 1970, Metropolitan plan making: An analysis of experiences with the preparation and evaluation of alternative land use and transportation plans (Regional Science Research Institute, Philadelphia, PA).

Hayashi, Y., 1983, A disaggregate model system for forecasting industrial location, Paper presented at the 8th Pacific Regional Science Conference, Tokyo.

Lee, K.S., 1982, A model of intra-urban employment location: An application to Bogota, Colombia, Journal of Urban Economics 12.

Lerman, S.R., 1977, Location, housing, automobile ownership, and mode to work: A joint choice model, Transportation Research Record, no. 610.

Lerman, S.R., 1983, Random utility models of spatial choice, Paper presented at the International Symposium on New Directions in Urban Modelling (University of Waterloo, Waterloo).

Nakamura, H., Y. Hayashi and K. Miyamoto, 1983, Land use/transportation analysis system for a metropolitan area, Transportation Research Record, no. 931.

Regional Science and Urban Economics 16 (1986) 145–160. North-Holland

URBAN GROWTH AND DECLINE IN A HIERARCHICAL SYSTEM

A Supply-Oriented Dynamic Approach

Roberto CAMAGNI

University L. Bocconi, Milan, Italy
University of Padua, Padua, Italy

Lidia DIAPPI

Politecnico of Milan, Milan, Italy

Giorgio LEONARDI

IASI-CNR, Rome, Italy

Received November 1984, final version received August 1985

In this paper a dynamic simulation model of urban growth and decline is presented, where innovation at the urban scale is crucial. The model is characterized in terms of supply conditions of the different functions or bundles of goods produced at the different ranks of the urban hierarchy. Urban dynamics depends on the form of the net location benefits curve, in relation to urban size and rank, and it is constrained by demand or market size conditions. Both urban growth and decline are linked to the presence of location benefits and the appearance of innovations or new production.

1. Introduction

Traditional and modern approaches to urban development exhibit a natural tendency to cluster around a few consolidated theoretical 'trajectories', that are highly characterized in terms of methodology or field of inquiry. This tendency, which on the one hand facilitates the self-perpetuation of traditions or schools of thought and the progressive sophistication of theories and models, is responsible on the other hand for the paucity of new approaches to the problem and, above all, for the weakness of linkages among the different approaches.

Thus, in spite of the evident interdependence of spatial phenomena, diverse aspects of urban growth have been studied 'per se' and insufficient efforts have been devoted to integrating the different theories into a unified or even 'eclectic' model [Wilson (1983b)].

We refer mainly to the central-place model of urban hierarchy, to the theoretical and empirical inquiries on optimal city size, to the export-base urban multiplier models and the spatial counterpart of the 'product life-cycle' and 'filter-down' hypotheses, to the theory of the inter-urban diffusion of innovation and the urban life-cycle model. All these approaches, taken separately, have almost exhausted their heuristic possibilities, but they may possibly supply new insights through the 'cross-fertilization' of their respective theoretical bases.

While based upon a dynamic approach and representing true methodological innovations, the most interesting recent models of urban growth, by Allen and Wilson [Allen and Sanglier (1981), Allen (1982), Wilson (1981), Diappi (1983)], suffer from a disease similar to the above in terms of underlying economic theory.

The fascinating routes that they open reside mainly in the fact that they highlight the possible alternative paths of development for the urban system, thus adding 'new perspective to historical geography' and re-evaluating such concepts, previously banished from scientific research, as 'historical accidents' or 'memory'. Now it becomes possible to 'chart the particular path which is chosen, with reason why it is the case...and to ask whether bundles of alternative paths can be grouped together in such a way that they constitute a *type* of city' [Wilson (1983a)].

But, from a theoretical point of view, these models do not identify any economic forces beyond the spatial interaction, the profit maximizing mechanism and the traditional, demand-oriented, urban multiplier effect to explain the structure and dynamic path of the urban hierarchy.

What is even more disappointing is that all recent approaches to urban development do not consider *economic innovation*, the truly dynamic element that, after Schumpeter, may be seen as the 'primum mobile' and the driving force in capitalist societies [Camagni (1984)]. Innovation does not only determine relative regional development, mainly in its form of technological progress in industry, but it also shapes relative urban growth, mainly through the creation of new producer or consumer services, the increasing sophistication of existing services, the improvement of tertiary functions within industry and their selective decentralization along the urban hierarchy [Andersson and Johansson (1984), Camagni and Cappellin (1984)].

The present paper was written to address this widespread dissatisfaction with respect to the present state of the art of urban analysis. It builds upon some basic ideas developed in a previous study [Camagni et al. (1984)] and presents a supply-oriented dynamic model that theoretically integrates three fundamental elements: innovation, urban hierarchy and spatial interaction. On the basis of this model, a computer simulation of the dynamics of an abstract urban system was run, in order to test the structural behaviour of the model and to ascertain the theoretical conditions for the emergence of an urban hierarchy.

2. The 'efficient' size of urban centers

The starting point of this work resides in the old standing question about the economic limits to urban growth. In fact, the idea of the existence of an 'optimal' city size, though fascinating, is contradicted by logical objections which limit its theoretical relevance and explain its poor empirical validation [Richardson (1972), Marelli (1981)].

This last statement is confirmed by the fact that in the real world urban decline is taking place not just in large primate urban centers, but also in medium size cities and even in small towns. Indeed, in the last decade the urban system of Northern Italy's Po plain has shown negative population growth rates not just for primary centers (7 out of 9), but also for secondary centers of 75,000 to 150,000 inhabitants (8 out of 19) and also for small centers of 20,000 to 75,000 inhabitants (27 out of 113) [Camagni et al. (1984)].

This phenomenon is neither explained by recent approaches to urban growth, nor by the city-life cycle model [Klaassen et al. (1981), Van den Berg and Klaassen (1981)]. But a new fruitful and more relevant hypothesis may be put forward: namely, the hypothesis that an *'efficient' city-size interval exists separately for each hierarchical city rank, associated with its specific economic functions.* In other words, for each economic function, characterized by a specific demand threshold and a minimum production size, a maximum city size also exists beyond which the urban location diseconomies overcome production benefits.

Let us assume that, for each *localized* economic function (F) or bundle of goods associated with a specific rank in the urban hierarchy, there exists:

(i) a minimum efficient production size ($A_0, A_1, \ldots$ in fig. 1) and a supply or average cost curve that becomes perfectly horizontal above that size (TAC), as is currently assumed by most industrial economists;

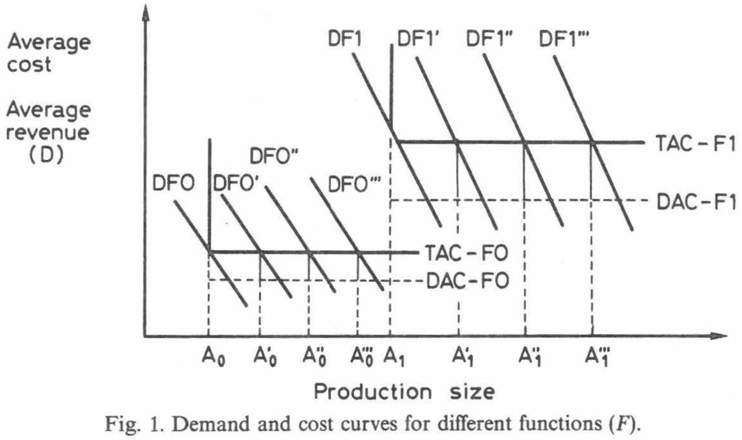


Fig. 1. Demand and cost curves for different functions (F).

(ii) a traditional (loeschian) demand curve (*D*) that is negatively shaped owing to the existence of spatial friction, for each income and population density level in the center and its surroundings; and consequently

(iii) a family of demand curves ($D', D'', \ldots$) as the demographic dimension of the single center increases (fig. 1). These curves define the equilibrium market production for each size of the center ($A'_0, A''_0, A'''_0, \ldots$), and the equilibrium average cost and revenue.

It is then possible for all functions ($F0, F1, \ldots$) ranging from the lower to the upper, to define a curve of average production benefit (*APB*), associated with the dimension of the *urban center* and defined by the 'mark-up' over equilibrium direct costs (*DAC*) (fig. 2). In this respect the city supplies both a spatially protected market that is not subject to distance decay, and broad availability and accessibility to qualified production factors.

Average profits may be assumed to increase as urban functions become of a higher order, due to (a) growing entry barriers, (b) decreasing elasticity of demand which allows extra profits to be gained in all market conditions far from the long-run equilibrium, and (c) increasing possibility of obtaining monopolistic revenues due to the use of scarce, qualified factors.

Moreover, we can directly compare this curve of average *production* benefits for each function with an Alonso type curve of average *location* costs (*ALC*), including land rent and congestion costs associated with urban size [Alonso (1971)] (fig. 2). Therefore, *for each economic function and each associated urban rank, it is possible to define a minimum and a maximum efficient city size*, which would increase with the level of the urban function and rank (A_0–A'_0 for the function and the center of rank $0, \ldots$).[1]

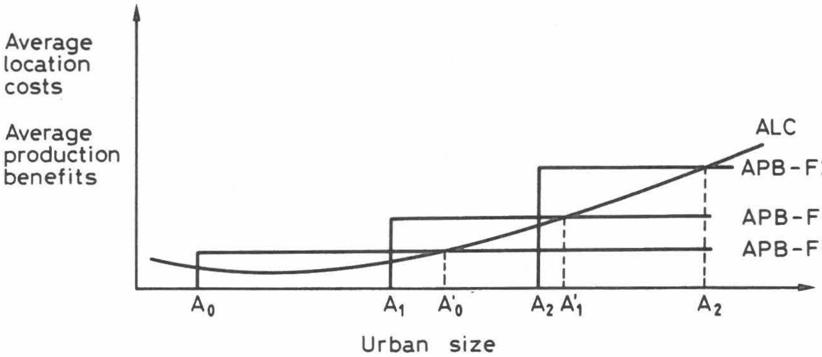


Fig. 2. Efficient urban size for different urban functions.

[1]All curves presented in figs. 1, 2 and 3 are 'average' and not 'marginal' curves as they refer to demand and supply conditions of the entire competitive *market* and not of a specific *firm*. In fig. 1, the spatial market of functions $1, 2, \ldots$ is illustrated; in figs. 2 and 3, on the other hand, a sort of aggregated urban land market is presented, with a supply curve which includes rent and congestion costs, and demand curves coming from the different urban functions.

As each center grows, it becomes potentially more suitable for the location of higher order functions, in terms of elements of both demand and supply. Lower order functions may be assumed to persist within higher order centers due to an intersectoral redistribution of the surplus attained by higher order functions.

In a dynamic setting, many elements may change the static picture presented hereto. Both growing per capita income which widens the market (holding population density constant) and fluctuations in income elasticity of demand and in relative physical productivity or terms-of-trade among the different functions may modify the efficiency interval for each city rank. These last two elements are particularly important from a theoretical point of view: in fact, they were shown to be directly linked to the k rate in Beckmann's model of urban hierarchy and to be the economic determinants of the shape of the urban rank-size distribution [Beckmann (1957), Beguin (1983)].

Another dynamic element is technical progress and in particular the application of microelectronics in industry, as it reduces minimum optimal production size in each sector or function. It may therefore generate, on the one hand, the spatial diffusion of higher order functions towards lower order centers (from A_1 to A_0 in fig. 3); on the other hand, it may create in larger centers a condition of oversize with reference to the maximum efficient urban size interval for their respective production (from A_1' to A_1'').

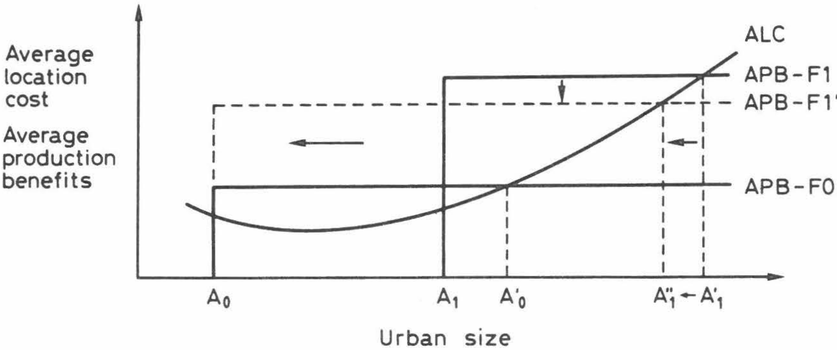


Fig. 3. Technical progress and efficient urban size.

3. Urban dynamics

Within each single interval, each center should grow according to its distance from an 'equilibrium' size where production benefits equal location costs; its path follows a logistic curve, which theoretically fits neatly into Wilson's 'unified' model of location and growth [Wilson (1983a)].

Population increase in this case may take place mainly through migration, from other centers and from outside the urban system considered.

This spatial interdependence describes that form of the dynamic behaviour of a city system which has been called *'constrained dynamics'*. This form 'refers to a system where the element of time plays an intrinsically important role in the evolution of state and/or control variables without, however, affecting the structure of the system itself' [Nijkamp and Schubert (1983)].

But another, more relevant, dynamic behaviour may be considered when 'the system configuration exhibits an incremental or integral change'. This behaviour is termed *'structural dynamics'*. In this case, innovation and bifurcation play the dominant conceptual role.

In our urban setting, each center's long-term growth possibilities are tied to its *ability to move to ever higher urban ranks, developing or attracting new and superior functions.* This ability is by no means mechanically attained, and does not spring directly from a simple market dimension, as in most traditional demand-side, export-base models.

Urban size, which is, however, a proxy not only for market size but also for presence of qualified production factors, is nothing but a necessary precondition for acquiring a new function. The real acquisition of a new function (n), once the size of the center has overcome its appearance threshold (A_n), depends upon the innovativeness of the private and public urban sectors and may be treated as a stochastic variable within the model.

As in Allen's model of urban dynamics [Allen (1982)], each center's growth path is subject to successive bifurcations which are linked to the appearance (within the correct intervals) of new economic functions as well as to the pace of general technical progress. The latter is responsible for sudden reductions in maximum efficient city size and for consequent urban decline in terms of population. Leaving aside the general spatial interaction among centers, the single center path may be described as in fig. 4.

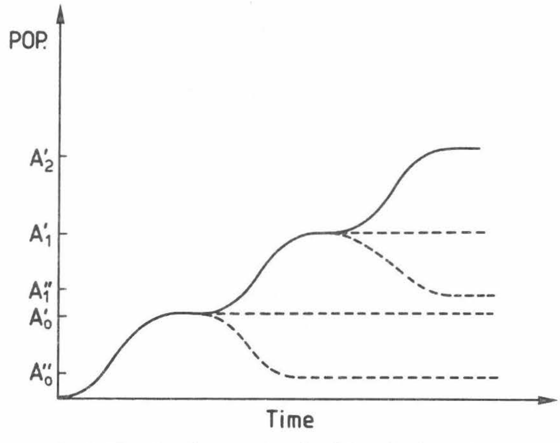


Fig. 4. The development path of the single center.

The probability of each center's entering a new phase of development by capturing a new function depends on many endogenous elements:

(a) the ability to overcome the minimum appearance threshold, which controls for the existence of appropriate production factors and of a minimum 'sheltered' local market;
(b) the possibility of a spillover or diffusion process from centers of a higher rank, located in close proximity;
(c) the diversification of local production, in terms of the presence of the entire range of activities or functions that characterize the single urban level [Chinitz (1961)]; indeed, a specialized oligopolistic urban structure is likely to be less innovative than a competitive, diversified one;
(d) the general situation of spatial competition with respect to the single new function. In fact, the existence of a sufficient market share for each center when it acquires a higher function is a condition for its persistence in the higher rank of centers. Through this competitive mechanism, demand is introduced into our supply-oriented model; nevertheless, differently from most of the existing models, it is conceived as a minimum threshold, not as the driving force in the dynamics of the urban system.

4. The model

The basic mechanisms of urban evolution may be expressed by two equations describing population growth at the single center j and the stochastic process of changing in rank.

Let $K=(k_1,\dots,k_j,\dots,k_n)$ represent the state vector of the rank of each center at time t. Then the differential equation of population growth in each center j of rank k within the interval of 'efficient' city size, a process we have labelled as 'constrained dynamics', may be defined as follows:

$$\dot{P}_j=P_j\left\{[B_{k_j}-C(P_j)]\left[a+m\sum_{i\neq j}P_i f(c_{ij})\right]-m\sum_{i\neq j}P_i[B_{k_i}-C(P_i)]f(c_{ji})\right\}, \qquad (1)$$

where P is population, a is the net migration rate from outside the urban system + the net natural growth rate, c is cost associated with distance, m is the interurban migration rate within the system, k denotes the urban rank and the associated economic function, C's are the average location costs, an increasing function of urban size, and B_k's are the average production benefits for function k in center j, as defined in fig. 2.

Moreover,

$$B_{k_j}=\bar{B}_{k_j} \quad \forall\, P>A_k, \quad \text{and}$$

$$f(c_{ij})=\exp(-q_{k_j}c_{ij}),$$

where q_k is an attraction or accessibility coefficient and A_k is the minimum appearance threshold for each function.

The equation describes essentially a logistic growth of population up to some limiting values which depend on the specific values of B_{k_j}, and therefore on the rank of center j.

A second equation defines the 'structural' dynamics of the urban hierarchy, which come into play when each center captures new, higher functions and consequently moves to higher ranks in the urban hierarchy.

In this respect, the process of urban growth and decline may be described through a stochastic process where the single center j of rank k constitutes the system. The state vector defines the probability of belonging to the rank k. The transition probability matrix is markovian and non-homogeneous, since the probabilities, defined as functions of population size in j, change with time.

If Π_k is the probability of belonging to rank k, its change in time may be defined as the sum of the probabilities of entry and exit, due to the gain or loss of functions k, $k-1$ and $k+1$,

$$\dot{\Pi}_k = \Pi_{k-1} \cdot GR_{k-1} + \Pi_{k+1} \cdot DC_{k+1} - \Pi_k(GR_k + DC_k). \tag{2}$$

GR_k is defined as the rate of change of the rank k of the city to the rank $(k+1)$ and expresses the ability of capturing new higher level functions.

The probability GR_k can be considered as the product of the following events:

- the overcoming of the appearance threshold A_{k+1} (in population terms) of the next higher level rank of economic activities,
- the existence of externalities or spillover effects coming from centers of higher order (EX), and
- the differentiation vs. specialization of local economic structure, representing a favourable condition for innovation and local creativeness; this element is expressed in terms of a Theil index of sectoral specialization (SP).

$$GR_k = g\{\exp[(P_j - A_{k+1})/A_{k+1}]\} \cdot hEX_k \cdot lSP_k, \quad \text{where}$$

$$EX_k = \sum_{k_i > k_j + 1} \sum_{i \neq j} P_i f(c_{ji}), \quad \text{and}$$

$$SP_k = \left[-\sum_n S_{jn} \ln S_{jn} \right]^{-1}.$$

S_{jn} is the economic dimension of the n sub-functions or sectors in center j, and h and l are normalizing factors. In its turn, DC_k is the rate of change from rank k of the city to rank $(k-1)$. This probability of losing function k (and to leave the corresponding urban rank k) depends on the overcoming of a demand constraint, given by the average market potential Φ of all centers which compete in the same function k,

$$DC_k = r \cdot [\Pi_{i \in i:(k_i \geq k_j)} \Phi_i^{w_{i(j)}}]/\Phi_j, \quad \text{where}$$

$$\Phi_j = \sum_i P_i f(c_{ij}), \quad \text{and}$$

$$w_{i(j)} = P_i \Big/ \sum_{i \in i:(k_i \geq k_j)} P_i.$$

5. The dynamic simulation

The temporal evolution of the stochastic process presented above has been studied through a random sampling simulation model based on a Monte Carlo procedure. This procedure is applied to the process of gain and loss of functions along the urban hierarchy; by this, the random character of the model and the importance attached to the innovation process are highly emphasized.

The simulation model allows us not only to analyse and compare the behaviour of the system under different parameter values, but also to evaluate the impact of different initial conditions upon the final asymptotic state of the system.

Particularly interesting initial conditions may be found (1) in some homogeneous spatial configuration describing an abstract early stage of urban development, and (2) in some theoretical equilibrium state of the hierarchical system, such as a 45° negatively sloped Zipf curve or a Christaller type spatial pattern of centers distribution.

To provide the basis for numerical experiments, an idealized geometrical zoning system is employed, in which the centers are arranged on a regular (triangular, quadratic or hexagonal) grid and where the distance among them is a parameter of the simulation (see appendix for details on the simulation procedures and parameters employed).

With respect to other similar dynamic simulation models, which are mainly concerned with sensitivity analysis and with the stability analysis of the asymptotic behaviour of the system, the major emphasis in our case is devoted to the simulation of different abstract processes bearing a precise theoretical interest.

The theoretical problems which the model has actually been used to deal with are the following:

(1) the effect of technical progress, represented by continuous shift in the

appearance thresholds of urban functions, on the spatial and size distribution of centers;

(2) the effect of different forms of the net benefits function ($B-C$: production benefits less location costs), and in particular the effect of different hypotheses concerning the average net returns to urban scale; in our specific case, we are referring to constant, decreasing and increasing *returns to urban rank*, as net benefits are steadily diminishing within each interval of 'efficient urban size' and only an innovation or a jump over a higher rank may increase them;

(3) the effect of different spatial deterrence parameters, with reference to both the general internal accessibility of the system and the relative spatial friction for different economic functions.

Two initial states of the system were chosen: an abstract state of uniform city rank distribution (with all, small-sized, centers randomly lying within the size interval of the second rank), and an equally random Zipf-type distribution of centers, ranging from the lowest ($<12{,}500$ inhabitants) to the highest ($>800{,}000$) of seven city ranks.

Starting from a set of parameter values taken from the real world experience of the Lombardy urban system, viz. birth, deaths and migration rates, general and relative spatial friction parameters, simulations were run in these alternative cases:

- high general spatial impedence, not presented here in detail, vs. a rapidly smoothing-down impedence with rising urban functions,
- constant, linearly increasing and exponentially increasing net returns to urban rank,
- fixed vs. variable appearance thresholds of the different functions, in order to simulate absence or presence of technical progress.

Nine cases are presented and discussed here in detail.

Case 1A. Homogeneous initial distribution and absence of technical progress; constant returns to urban scale.

Case 1B. The same conditions as before, but linearly increasing returns.

Case 1C. The same conditions as before, but exponentially increasing returns.

Case 2A. Zipf-type initial distribution and absence of technical progress; constant returns to urban scale.

Case 2B. The same conditions as before, but linearly increasing returns.

Case 2C. The same conditions as before, but exponentially increasing returns.

Case 3A. Zipf-type initial distribution and diminishing appearance thresholds (50% in the first 50 years); constant returns to scale.

Case 3B. The same conditions as before, but linearly increasing returns.

Case 3C. The same conditions as before but exponentially increasing returns.

6. Main results

The main results of the simulation may be summarized as follows.

(1) The simulation model shows a strong internal consistency, due to the high interdependence of its parts (eq. 1), and a strong stability in time. Indeed two hundred years were necessary to create the entire urban hierarchy in the case of homogeneous initial distribution (fig. 5).

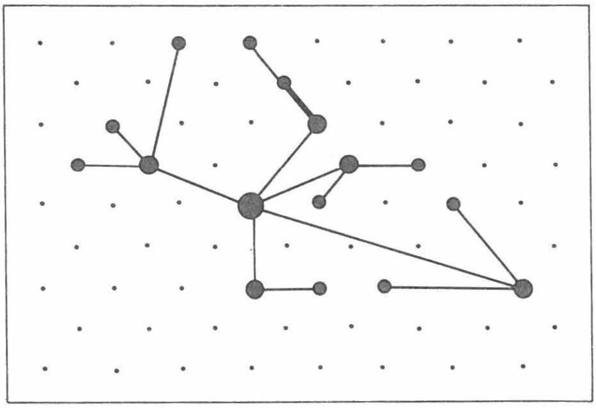


Fig. 5. Final spatial structure of centers (rank 5, 6, 7) when the initial state of the system is homogeneous.

(2) Higher probabilities of decline are found at the periphery of the system, where it is difficult to overcome the minimum demand threshold.

(3) A general condition for the creation of an urban hierarchy seems to reside in the presence of increasing returns to urban scale (or rank). In fact, starting from a homogeneous initial condition, a very flat hierarchy organized over only four ranks was apparent after 200 runs in the constant returns hypothesis (tables 1 and 2); moreover, in the Zipf-type initial distribution of centers, the hierarchical structure is hardly maintained under constant returns (case 2A), as it is shown by the flattening of the slope of the Zipf curve (from 1.075 at $t=0$ to 1.028 at $t=100$), the diminishing importance of the prime center (from 20% to 18% of total population) and the general shift towards lower order centers. This is in our view one of the most

Table 1

Rank-size distribution of centers in the final state.[a]

Cases	Year	Total population	(a) Intercept	(b) Slope	R^2
1A	200	6.976.000	2.46	−0.348	0.750
1B	200	11.388.000	3.34	−0.926	0.930
1C	200	17.942.000	3.57	−0.954	0.918
Initial	0	6.282.000	3.18	−1.075	0.985
2A	100	7.126.000	3.19	−1.028	0.989
2B	100	7.736.000	3.27	−1.081	0.986
2C	100	18.043.000	3.80	−1.189	0.937
3A	100	6.888.000	3.22	−1.054	0.987
3B	100	7.432.000	3.31	−1.101	0.978
3C	100	15.769.000	3.71	−1.138	0.851

[a] $\ln Pop = \log a - b \ln(\text{rank } k)$.

Table 2

Frequency of centers in each rank of the urban hierarchy.

		Rank						
Cases	Year	1	2	3	4	5	6	7
1A	200	5	13	42	12	0	0	0
1B	200	15	21	9	17	7	3	0
1C	200	10	18	12	16	10	5	1
Initial	0	26	21	14	5	3	2	1
2A	100	38	13	10	6	4	0	1
2B	100	37	14	9	7	3	1	1
2C	100	13	20	14	10	6	6	3
3A	100	10	27	13	12	6	3	1
3B	100	10	27	13	9	7	4	2
3C	100	8	5	19	16	6	12	6

interesting results of the simulation model, as it adds to Christaller's key concepts of 'demand threshold' and 'range' a further economic condition for hierarchization of centers, along theoretical lines similar to those recently highlighted by Beguin (1983).

(4) The absence of a high generalized spatial impedence was proved to be another, expected, condition for the creation of an urban hierarchy. In fact the simulations run under homogeneous spatial deterrence functions for the different rank-dependent bundles, omitted here, showed a marked difficulty of the higher centers to stabilize and even reach a sufficient market and population size.

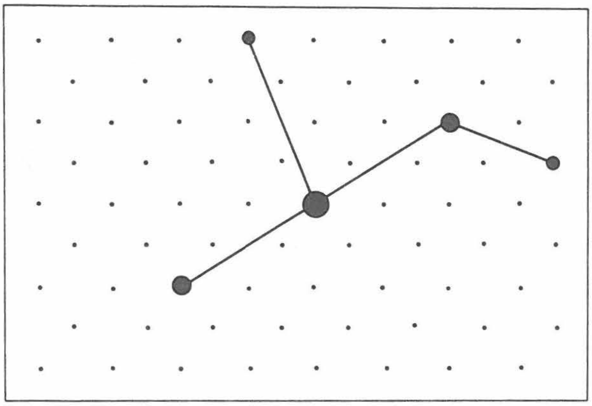


Fig. 6. The spatial structure of the urban hierarchy (rank 5, 6, 7) in case 2B, i.e., Zipf-type initial distribution, no technical progress, and linearly increasing returns to urban scale.

(5) In the Zipf-type initial distribution of centers, the hypothesis of linearly increasing returns with urban rank (case 2B) allows the persistence of the urban hierarchy in its initial shape and spatial pattern (fig. 6): a sort of steady-state in which population increases due to natural growth and migration from outside the system, leaving the relative size of centers almost untouched.

(6) The hypothesis of exponentially increasing returns, which strongly favours the innovative centers of the highest ranks, creates a steeper distribution of centers and a wider number of top cities (3 vs. 1 in the 2C case: see tables 1 and 2). These cities are not necessarily the 'historical' ones: an innovative center with a favourable position in the entire system may overcome initially higher ranked cities in this case (fig. 7).

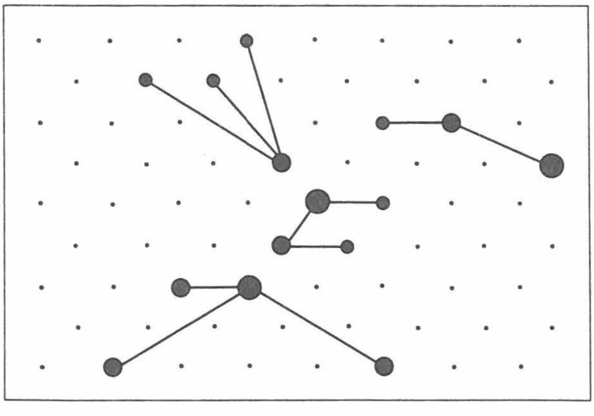


Fig. 7. The spatial structure of the urban hierarchy (rank 5, 6, 7) in case 2C, i.e., Zipf-type initial distribution, no technical progress, and exponentially increasing returns to urban scale.

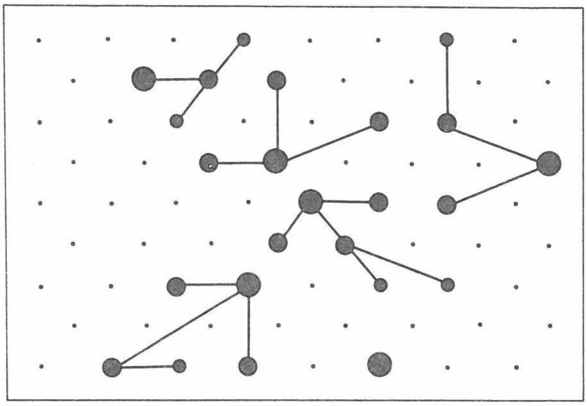


Fig. 8. The spatial structure of the urban hierarchy (rank 5, 6, 7) in case 3C, i.e., Zipf-type initial distribution, no technical progress, and exponentially increasing returns to urban scale.

(7) The previous tendency towards a policentric urban structure is strongly emphasized in case technical progress was taken into account: six centers are found in the sixth and seventh rank in the 3B case (against two in the 2B case) and 18 in the 3C case (against nine in the 2C case) (see fig. 8). In fact, due to the shifting down of appearance thresholds of the different functions, higher functions may be easily captured by smaller centers and the 'prime' urban role almost disappears: the biggest city accounts for only 11% and 5% of total population in the 3B and 3C case respectively. Only in the constant returns case (3A) a traditional hierarchy persists, though at lower population levels, as the shifting down of the appearance thresholds goes in parallel with the shifting down of the entire urban system, as seen before.

7. Conclusions

In this paper a dynamic simulation model of urban growth and decline is presented, where innovation at the urban scale is crucial. The model is deeply characterized in terms of supply conditions of the different functions or bundles of goods produced at the different ranks of the urban hierarchy. Urban dynamics depends on the form of the net location benefits curve, in relation to urban size and rank, and it is constrained by demand or market size conditions.

Both urban growth and decline are linked to two kinds of elements: the presence of positive location benefits in the actual production activities, and the appearance of innovations or new production, which generate bifurcations in the historical path of the single centers.

Different spatial configurations of the urban hierarchy are the outcome of different hypotheses concerning the initial state of the system, the presence of technical progress and the presence of increasing net returns to urban scale. The latter condition is proved to be crucial for the formation of an urban hierarchy.

Appendix: Parameter values of the simulation model

Number of centers: 72.
Number of ranks: 7.
Distance among centers: 20.
Minimum appearance thresholds for the different functions:
12.5 25 50 100 200 400 800 (thousands of inhabitants of the center).
Maximum efficient city size:
30 60 120 240 480 880 1600 (thousands of inhabitants of the center).
Frequency of centers in each rank in the initial state:
– homogeneous case: 0 1.00 0 0 0 0 0,
– Zipf-type case: 0.39 0.25 0.18 0.11 0.04 0.02 0.01.

$\Delta t = 1$ year.
b = birth rate = 0.003.
d = death + outmigration rate = 0.003.
a = intra-system migration rate = 0.08.
q = decay function coefficient for migration movements = 1/60.
Transportation cost coefficients for each function:
1/20 1/40 1/80 1/160 1/320 1/640 1/1280.
g = growth probability coefficient = 0.03.
r = decline probability coefficient = 0.03.

References

Allen, P.M., 1982, Order by fluctuation and the urban system, in: A.A.V.V., ed., New quantitative techniques for economic analysis (Academic Press, London).
Allen, P.M. and M. Sanglier, 1981, Urban evolution, self-organization and decisionmaking, Environment and Planning A 13, 167–183.
Alonso, W., 1971, The economics of urban size, Papers and Proceedings of the Regional Science Association XXVI.
Andersson, A.E. and B. Johansson, 1984, Knowledge intensity and product cycles in metropolitan regions, Contributions to the Metropolitan Study 8, WP-84-13 (IIASA, Laxenburg).
Beckmann, M.J., 1957, City hierarchies and the distribution of city size, Economic Development and Cultural Change 6, 243–248.
Beguin, H., 1983, The shape of city-size distribution in a central place system, Paper presented at the Twenty-Third European Congress, RSA, Poitiers, 30 Aug.–2 Sept.
Camagni, R.P., 1984, Les modèles de restructuration économique des régions européennes pendant les années 70, in: Ph. Aydalot, ed., Crise et espace (Economica, Paris).

Camagni, R. and R. Cappelin, 1984, Structural change and productivity growth in the European regions, in: N. Groes and K. Peschel, eds., Regional Research in an Interregional Perspective (Florentz, München).
Camagni, R.P., F. Curti and M.C. Gibelli, 1984, Ciclo urbano: Le citta' tra sviluppo e declino, Scienze Regionali 3 (Franco Angeli, Milano).
Chinitz, B., 1961, Contrasts in agglomeration: New York and Pittsburg, Papers of the American Economic Review, May.
Diappi, L., 1983, Processi di autoorganizzazione nei sistemi urbani, Proceedings of the Third Italian Regional Science Conference, Firenze, Nov.
Klaassen, L.H., W.T.M. Molle and J.H.P. Paelinck, eds., 1981, Urban Europe: A study of growth and decline (Pergamon, Oxford).
Marelli, E., 1981, Optimal city size, the productivity of cities and urban production function, Sistemi Urbani 3, nos. 1–2, 149–163.
Nijkamp, P. and U. Schubert, 1983, Structural change in urban systems, Contributions to the Metropolitan Study no. 5 (IIASA, Laxenburg).
Parr, J.B., 1981, Temporal change in a central-place system, Environment and Planning A 13, 97–118.
Richardson, H.V., 1972, Optimality in city size, Systems of cities and urban policy: A sceptic's view, Urban Studies 9, 29–47.
Van den Berg, L. and L.H. Klaassen, 1981, The contagiousness of urban decline (N.E.I., Rotterdam).
Wilson, A.G., 1981, Catastrophe theory and bifurcation: Applications to urban and regional systems (Croom Helm, London).
Wilson, A.G., 1983a, Location theory: A unified approach, Working paper no. 355 (School of Geography, University of Leeds, Leeds) May.
Wilson, A.G., 1983b, Making urban models more realistic: Some strategies for future research, Working paper no. 358 (School of Geography, University of Leeds, Leeds) Nov.