STUDIES IN REGIONAL SCIENCE AND URBAN ECONOMICS 20

ADVANCES IN SPATIAL THEORY AND DYNAMICS

Å. E. ANDERSSON
D. F. BATTEN
B. JOHANSSON
P. NIJKAMP
editors

North-Holland

Studies in Regional Science and Urban Economics

Series Editors

ÅKE E. ANDERSSON
WALTER ISARD
PETER NIJKAMP

Volume 20

NORTH-HOLLAND – AMSTERDAM • NEW YORK • OXFORD • TOKYO

ADVANCES IN SPATIAL THEORY AND DYNAMICS

Advances in Spatial Theory and Dynamics

Editors

ÅKE E. ANDERSSON
Institute for Future Studies, Stockholm
University of Umeå
Sweden

DAVID F. BATTEN
BÖRJE JOHANSSON
University of Umeå
and Center for Regional Science Research
Sweden

and

PETER NIJKAMP
Free University
Amsterdam, The Netherlands

1989

NORTH-HOLLAND – AMSTERDAM • NEW YORK • OXFORD • TOKYO

ISBN: 0 444 87357 0

Published by:

ELSEVIER SCIENCE PUBLISHERS B.V.
P.O. Box 1991
1000 BZ Amsterdam
The Netherlands

Sole distributors for the U.S.A. and Canada:

ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
655 Avenue of the Americas
New York, N.Y. 10010
U.S.A.

PRINTED IN THE NETHERLANDS

GIORGIO LEONARDI, 1942-1986

IN MEMORIAM

INTRODUCTION TO THE SERIES

Regional Science and Urban Economics are two interrelated fields of research that have developed very rapidly in the last three decades. The main theoretical foundation of these fields comes from economics but in recent years the interdisciplinary character has become more pronounced. The editors desire to have the interdisciplinary character of regional sciences as well as the development of spatial aspects of theoretical economics fully reflected in this book series. Material presented in this book series will fall in three different groups:

- interdisciplinary textbooks at the advanced level,
- monographs reflecting theoretical or applied work in spatial analysis,
- proceedings reflecting advancement of the frontiers of regional science and urban economics.

In order to ensure homogeneity in this interdisciplinary field, books published in this series will:

- be theoretically oriented, i.e. analyse problems with a large degree of generality,
- employ formal methods from mathematics, econometrics, operations research and related fields, and
- focus on immediate or potential uses for regional and urban forecasting, planning and policy.

Åke E. Andersson
Walter Isard
Peter Nijkamp

THE INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS

is a nongovernmental research institution, bringing together scientists from around the world to work on problems of common concern. Situated in Laxenburg, Austria, IIASA was founded in October 1972 by the academies of science and equivalent organizations of twelve countries. Its founders gave IIASA a unique position outside national, disciplinary, and institutional boundaries so that it might take the broadest possible view in pursuing its objectives:

To promote international cooperation in solving problems arising from social, economic, technological, and environmental change

To create a network of institutions in the national member organization countries and elsewhere for joint scientific research

To develop and formalize systems analysis and the sciences contributing to it, and promote the use of analytical techniques needed to evaluate and address complex problems

To inform policy makers and decision makers of how to apply the Institute's methods to such problems

The Institute now has national member organizations in the following countries:

Austria
The Austrian Academy of Sciences

Bulgaria
The National Committee for Applied Systems Analysis and Management

Canada
The Canadian Committee for IIASA

Czechoslovakia
The Committee for IIASA of the Czechoslovak Socialist Republic

Finland
The Finnish Committee for IIASA

France
The French Association for the Development of Systems Analysis

German Democratic Republic
The Academy of Sciences of the German Democratic Republic

Federal Republic of Germany
Association for the Advancement of IIASA

Hungary
The Hungarian Committee for Applied Systems Analysis

Italy
The National Research Council

Japan
The Japan Committee for IIASA

Netherlands
The Foundation IIASA–Netherlands

Poland
The Polish Academy of Sciences

Sweden
The Swedish Council for Planning and Coordination of Research

Union of Soviet Socialist Republics
The Academy of Sciences of the Union of Soviet Socialist Republics

United States of America
The American Academy of Arts and Sciences

Preface

This volume originates from two meetings, set apart in time but closely connected by continuing collaborative efforts between researchers in an international network. The first of these meetings took place at The International Institute for Applied Systems Analysis (IIASA) in October 1984, organized by IIASA:s Regional Issues Project under the title "Dynamic Analysis of Spatial Development". About half of the papers in this volume were presented at that meeting. These contributions have been elaborated and revised considerably during the preparation of the volume, and can now be regarded as mature papers embracing the frontiers of spatial and economic dynamics..

Another set of contributions was presented during the European Summer Institute in Regional Science held at the University of Umeå in June 1986. The Summer Institute was organized by CERUM (Centre for Regional Science at Umeå University) in collaboration with the Departments of Economics and Geography at the same university. The contributions have been drawn from the sessions on technological change, nonlinear dynamics in spatial networks and infrastructure development. This is reflected in the three parts of the volume (1) Competition, specialization and technological change, (2) Spatial interaction, (3) Urban and regional infrastructure.

In essence there is an unbroken line between the first and second meeting, with a joint focus on slow adjustment processes governing the change of urban and internodal network infrastructure. Many of these studies have been supported by the Swedish Council for Building Research (BFR). A related issue is the interdependence between economic and spatial processes operating on different time scales. Studies in this field were initiated in the beginning of the 1980's within the IIASA-project "Nested Dynamics of Metropolitan Processes and Policies". Many of the ideas in this project originated from collaboration with our friend and colleague Giorgio Leonardi, whose passing soon after the Summer Institute in 1986 was a great loss to so many of us. This volume is dedicated to his memory.

Acknowledgements

The editorial work on this volume has been undertaken by the Centre for Regional Science Research (CERUM) at the University of Umeå. Jenny Wundersitz coordinated all the editorial tasks at CERUM and Ingrid Lindqvist prepared and revised the manuscript. The publications department at IIASA has finalized the illustrations. Financial support came from Jan Wallander's Foundation, Gösta Skoglund's International Fund and the Swedish Council for Building Research. The latter foundation has also supported several of the contributions in this volume (Grant 850205-06).

The editors are sincerely grateful to all of the above for their important contributions.

Åke E. Andersson David Batten Börje Johansson

Umeå University
July 1988

CENTRE FOR REGIONAL SCIENCE RESEARCH (CERUM)

CERUM is a research organisation at Umeå University in Sweden. Its basic objective is to participate in and contribute to the development of regional science in a national and international context. In view of this goal, the centre coordinates nation-wide research programmes; it promotes studies focussing on spatial aspects of regional, national and international social, political and economic systems. The research emphasizes the development and assessment of theories as well as pertinent models and applied methods of research. The centre also hosts the Secretariate for Systems Analysis which aims at facilitating the interaction between Swedish researchers as well as research organisations and the International Institute for Applied Systems Analysis (IIASA) in Austria. In this effort the secretariate is guided by the Swedish member organisation of IIASA, i.e., The Swedish Council for Planning and Coordination of Research.

Research Orientation

The research tradition within regional science is multidisciplinary.The discipline has developed as a result of contributions from many scientific fields such as civil engineering, demography, geography, economics, sociology, political science, urban and regional planning, applied mathematics and systems analysis. CERUM's task is to continue with this tradition so as to utilize and combine contributions from the social, natural and engineering sciences in the study of spatial and regional development.

CERUM has two in-house series of publications: (1) Working Papers from CERUM, and (2) CERUM Reports. The centre has a national board with current members from *the universities of Lund, Luleå, Umeå, Uppsala, Sundsvall-Härnösand, Östersund; the Royal Institute of Technology, the Association of Swedish Municipalities, the Expert Group of Regional Development Research of the Ministry of Industry, the Institute for Future Studies, the National Swedish Board of Industry, and the National Swedish Board for Technical Development.*

CERUM

Centre for Regional Science Research

CONTENTS

PART C: URBAN AND REGIONAL INFRASTRUCTURE

Advances in Spatial Theory and Dynamics
Å.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 1

Progress in Spatial Theory and Dynamics: A Prefatory Review

D.F. Batten and P. Nijkamp

1. INTRODUCTION

During the late sixties and seventies, the youthful field of regional science moved gradually towards a more mature phase as the interdisciplinary scholarship of economists, geographers, mathematicians, engineers and others displayed some welcome signs of synthesis and consensus. The basic spatial problems - those associated with location, market areas, land use, trade and regional development - were painted clearly and broadly by Walter Isard and others in the fifties (see, for example, Isard, 1956). Important progress has been made in tackling these well-posed problems, perhaps most notably in the fusing of location theory with spatial interaction models, transportation planning and regional economic analysis.

But new problems had already emerged before these old ones could bask in welcome solution. The late seventies and eighties have seen pronounced slowdowns in the economic or population growth rates of many places - be they cities, regions or nations. Widespread processes of economic and physical obsolescence are becoming relatively severe in some of these places. Quite often we have witnessed abrupt changes to the smooth urbanization trends experienced in earlier decades.

These sudden discontinuities and structural economic changes prompt a need for new concepts and new tools which can probe beyond the traditional optimization and lifecycle theories familiar to regional scientists. In short, technological changes and the march towards internationalization may be expected to have a profoundly different impact on spatial patterns of human activity than we have witnessed in earlier decades.

Spatial dynamics can be observed in all countries of the world. In some cases, regions and cities display a smooth transition pattern, while in others sudden changes take place. The Silicon valley development pattern, the rapid expansion of the Shinkansen Region and the Greater Boston Area, and the growth of metropolitan areas in many developing countries reflect a transition that may be denoted as structural dynamics and which may exhibit unstable systems behaviour. Unstable behaviour may lead to sudden oscillations in the prosperity enjoyed by particular places, and considerable uncertainty with regard to the future.

But perhaps there are even more convincing reasons for a necessary review of our progress in the field of spatial dynamics. It seems probable that we have recently entered a structural transition of global dimensions, which might be described as a Logistical Revolution (see Andersson, 1985); associated with the growth of information processing

and communication capacity as well as an expansion of the knowledge base. This development goes hand-in-hand with further improvements to the transportation system, especially to the structure and operation of the air transport network. Improvements to telecommunication and air transport capacity are jointly increasing the discrete network character of the world economy. Contiguity of places and regions is thus becoming less significant. Our traditional theories associated with central place hierarchies may be seen as less convincing. Instead we must concern ourselves with changes occuring across multiple layers of networks (i.e. people, commodities, money, information) from one hub or node to another.

The above discussion serves to emphasize the need for an even more comprehensive theory of social and economic development. Not only should such a theory embrace both the time and space dimensions, but it must also recognize the dynamic interplay of forces transmitted across several different levels and layers of spatial networks. Knowledge capacity is rapidly becoming a fundamental nodal parameter when we study networks of information exchange. Speed and certainty of communication is beginning to affect our choice of telecommunication device, just as speed of movement has traditionally played an important part in the modal choices made by travellers.

Despite relevant partial contributions, a unifying theory for analysing evolutionary patterns over space is lacking. It has been suggested by several authors that technological progress may be an important factor behind spatial development patterns, though only a few operational attempts have been made to include innovation as an *endogenous* impulse in spatial growth patterns. In order to shed more light on the intriguing role of innovation in spatial development patterns, the next section will be devoted to a concise discussion of long wave theories and innovations, and to their relevance for spatial development fluctuations. For these purposes, both *external* and *internal* factors affecting spatial dynamics will be taken into account.

2. SPATIAL DYNAMICS

2.1 External Determinants

Spatial economic systems have always been characterized by a state of flux. This dynamics may to a certain extent be ascribed to drastic changes in the environment *outside* the urban system leading to profound changes inside the system itself. For instance, the rise of oil prices in the seventies had a great impact on urban transportation systems and urban residential patterns (see Beaumont and Keys, 1982). A revival of interest in structural economic changes has therefore emerged, not only in the economic sense of innovation patterns but also in the geographical sense of reorientation of cities and regions (see e.g. Johansson and Nijkamp, 1987).

For many decades, economic fluctuations, long wave patterns and spatial dynamics have always commanded a great deal of attention by economic historians (cf. Adelman, 1965; and Schumpeter, 1939), but the emergence of the current economic recession and its inherent future uncertainty has stimulated a new interest in structural dynamics of economic systems (including *inter alia* such issues as industrial perturbations, unbalanced growth, disequilibrium analysis, international and geographical equity, and multi-actor conflicts (see e.g. Olson, 1982). In this respect, Kondratieff's theory of long cycles has stimulated new reflections and scientific debates (see, for instance Freeman et al., 1982, Mandel, 1980; and Mensch, 1979). Kondratieff's original theory distinguished five stages in the long-run cyclical pattern of a free enterprise economy: take-off, rapid growth, maturation, saturation and decline. The real existence of such long-term fluctuations is difficult to demonstrate due to lack of historical data. In general, only price

data have been used to test the long-wave hypothesis, although fortunately in recent years new efforts have been made to provide a more substantial empirical foundation for the long-wave hypothesis by means of industrial innovation and R&D data (see Kleinknecht, 1986). Several other cycles of shorter duration (e.g. Kuznets and Juglar cycles) may also be of interest.

2.2 Internal Determinants

An urban or regional economy may also display fluctuations that are internal to the system, caused *inter alia* by social, demographic, political, or economic forces. Such developments are particularly related to innovations in the industrial sectors, either *basic* innovations (leading to new products, new firms or even new industrial sectors) or *process* innovations (leading to new industrial processes in existing sectors). Basic innovations are assumed to take place periodically and to cluster, leading to economic fluctuations. In this respect it is usually assumed that after a period of growth a period of saturation may take place, leading to a recession. Such growth processes are often described by means of a logistic (S-shaped) curve characterized by the following phases: introduction, growth, maturity, saturation and decline (see, e.g. Batten, 1982).

Apart from innovations *per se*, the filtering and diffusion processes by which new inventions spread and evolve warrant our attention. For instance, although new innovations may emerge in nodal centres, in the long-run these innovations or their offspring may be observed elsewhere. It is here that the spatial version of product cycle theory can provide useful insights. Especially during a phase of saturation and decline, basic innovations and radical technological changes may be effective vehicles for a developing economy.

When we consider supply and demand perspectives simultaneously, the appropriate combination of R&D capital, production equipment, public overhead capital (or network infrastructure) and new markets is a necessary ingredient for creating radical technological changes (see Schmookler, 1966). Such changes are essentially the catalytic factors behind the process of structural economic development in various places.

2.3 The Importance of Network Infrastructure

The presence of satisfactory urban or regional infrastructure is thus a necessary condition to foster a breeding place for new activities (Rosenberg, 1976). This requires, in general, favourable educational facilities, communication possibilities, market entrance, good environmental conditions and agglomeration favouring innovative activities. In recent studies, dense import networks entering a node have been recognized as an innovation promoting system property (Johansson and Westin, 1987). The above may also explain why monopoly situations and industrial concentrations (including patent systems) often have greater technological and innovative opportunities. There is some empirical evidence to suggest that only a limited number of industrial sectors account for the majority of innovations (such as electronics, petrochemicals and aircraft) although in various cases small firms may be a source of major innovations (see Rothwell, 1979). This also implies that economic sectoral adjustment and spatial fluctuations may go hand in hand when perceived in terms of nonlinear relative dynamics (see e.g., Batten, 1985).

Another important question asks whether or not city size is a decisive factor for economic fluctuation and structural adjustment. In the debate on geographical concentration and specialisation, it is often argued that city size favours innovative propensity (cf. Alonso, 1971; Bluestone and Harrison, 1982; Jacobs 1977; Pred, 1966;

and Richardson, 1973). More recently we may observe that the innovative potential in the U.S. - which was traditionally concentrated in large urban agglomerations - may now be declining in the largest urban concentrations (see Markusen et al., 1986; Norton and Rees, 1979; and Sveikauskas, 1979).

Part of the explanation for this trend reversal may lie in the global transition towards a network economy. Growth seems to be proportional to size in a central place city but not necessarily in a network city. Network cities have accounted for most of the recent additions to the urban array, notably the fast-growing industrial newcomers (see Hohenberg and Lees, 1985). This argument is summed up in Figure 1, a simple schema that builds on the work of Robson (1973). Evolutionary models drawn from various scientific disciplines may also be useful for the analysis of innovation and spatial diffusion processes (see, e.g. Batten et al., 1987; Nelson and Winter, 1982). In this tradition one should observe that urban regions have vintage properties. This means that large and previously prosperous cities and metropolitan nodes may start to decline because of obsolescence. Cyclic patterns result when infrastructure investments bring about renewal of the economic system.

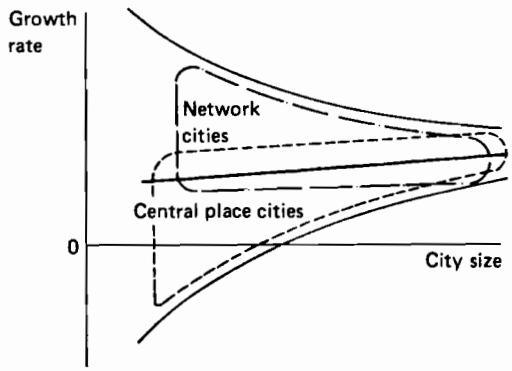


Figure 1 The relationship of growth and size in two urban systems.

3. AN ILLUSTRATION OF SPATIAL FLUCTUATION

3.1 Capital Formation in a Spatial System

The evolution of a spatial system may demonstrate fluctuations, unbalanced growth or perturbations. In the present section, a more formal approach to spatial long-term fluctuations will be presented (based on the previous section). First, the main driving forces of a spatial system will be described by means of a simplified arrow diagram (see Figure 2). The assumption made here is that R&D capital is assumed to incorporate knowledge and communication technology. Various production factors may thus exert an impact on spatial dynamics, as reflected in the impact model of Figure 1. For the moment, the diffusion processes of innovations will be excluded.

A simple mathematical representation of the driving forces of such a system can be found in Nijkamp (1986). This simplified model was based on a *quasi-production*

function (including productive capital, infrastructure and R&D capital as factors). The dynamics of the system was described by motion equations for productive investments, infrastructure investments and R&D investments. Several constraints were also included to control congestion effects and consumption rates. Equilibrium solutions were obtained using optimal control theory.

It is evident that if we focus on *qualitative* changes in a non-linear dynamic system, sudden jumps and discontinuities may emerge (see e.g., Allen and Sanglier, 1979; Andersson and Kuenne, 1986; Batten, 1982; Dendrinos, 1981; Isard and Liossatos, 1979; Mees, 1975; and Wilson, 1981). In the following section the issue of *non-linear* dynamics will be demonstrated using a well known equation in difference form.

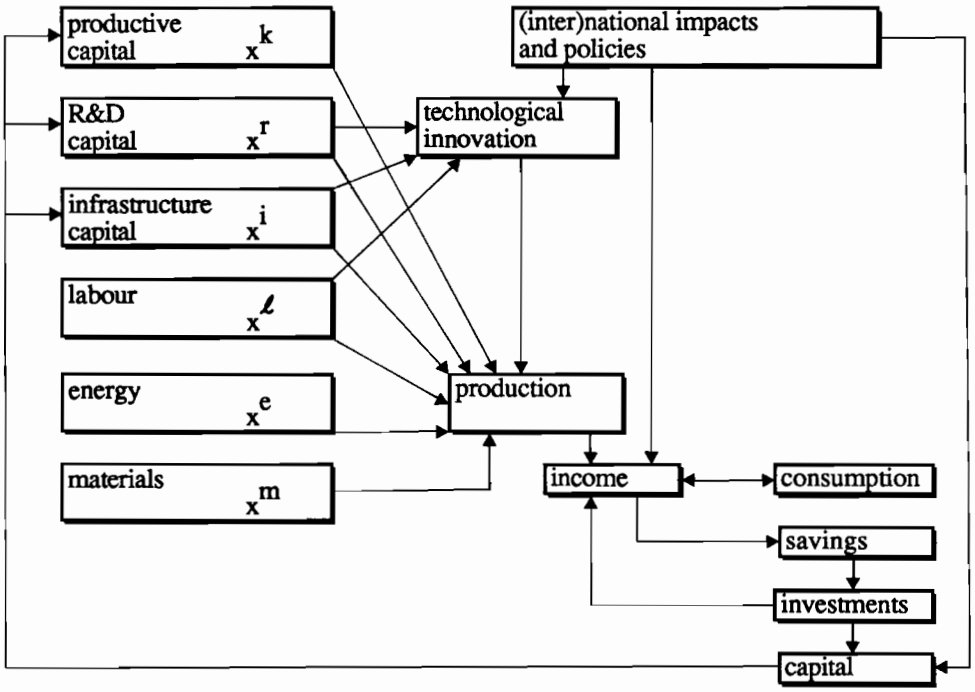


Figure 2 The process of capital formation

3.2 A Simple Model for Generating Spatial Fluctuations

Consider a (closed) urban or regional economy characterized by a 'generalized' production function including productive capital (x^k), labour (x^ℓ), energy (x^e), materials (x^m) network infrastructure (x^i) and R&D activities (x^r) as arguments. The first four components are often found in KLEM production functions dealing with substitution effects between capital, labour, energy and materials (see for instance Lesuis et.al., 1980). The fifth component indicates the necessary public overhead capital needed as a complement to private productive capital, along the lines suggested by Hirschman (1958), in order to facilitate spatial interaction and so achieve a balanced growth strategy. The inclusion of this infrastructure component led to the notion of the quasi-production

function in recent literature (see Nijkamp, 1984). Finally the sixth component is reflecting the innovation effects due to R&D investments (including knowledge capacity and communication) in the spatial system.

Hence, the following generalized production function may be assumed:

$$y = f(x^k, x^{\ell}, x^e, x^m, x^i, x^r), \tag{1}$$

where y is the volume of production (in terms of market share). The parameters of the location-specific production technology depend on the general state of the technology worldwide and on specific agglomeration factors at the urban or regional level.

If a normal Cobb-Douglas specification is assumed, one may rewrite (1) as follows:

$$y = \alpha(x^k)^{\beta} (x^{\ell})^{\gamma} (x^e)^{\delta} (x^m)^{\varepsilon} (x^i)^{\xi} (x^r)^{\eta}, \tag{2}$$

where the parameters $\beta,\ldots,\eta$ reflect the production elasticities concerned. These elasticities are assumed to be positive in the range $(y^{min}), (y^{max})$. Below a certain minimum threshold level, y^{min}, the economy may be too small for agglomeration advantages to accrue, and thus a marginal increase in one of the production factors may have a zero impact on the total production volume. In other words a city needs a minimum endowment of production factors before reaching a self-sustained growth. Furthermore, beyond a certain maximum capacity level of urban size, bottlenecks and capacity tensions (due to a high concentration of capital) may cause a negative marginal product of some of the production factors (e.g., productive capital, R&D).

Shifts in the urban or regional production volume over a certain period of time may be written in difference form as:

$$\Delta y_t = (\beta k_t + \gamma \ell_t + \delta e_t + \varepsilon m_t + \xi i_t + \eta r_t) y_{t-1} \tag{3}$$

with

$$\Delta y_t = y_t - y_{t-1} \tag{4}$$

and

$$h_t = \frac{x_t^h - x_{t-1}^h}{x_t^h - 1}, \quad h = k, \ell, e, m, i, r \tag{5}$$

Thus the arguments of (5) are written as relative changes in terms of the original variables.

Within the range (y^{min}, y^{max}), the urban system will exhibit a noncyclical growth pattern characterized by stable behaviour and self-sustained growth. This self-sustained growth path may be curtailed from two causes:

- external factors: scarcity of production factors or lack of demand
- internal factors: emergence of congestion effects leading to negative marginal products.

External factors imply that the system will move towards an upper limit set by the constraint concerned. Internal factors may lead to perturbations and qualitative changes in systemic behaviour. Consider, for instance, a congestion effect caused by the exessive concentration of capital in an urban agglomeration. Then each additional increase in productive capital will have a negative impact on the urban production level. In other words, beyond the capacity limit, y^{max}, an auxiliary relationship reflecting a negative marginal capital product may be assumed:

$$\beta_t = \hat{\beta}(y^{max} - \kappa y_{t-1})/y^{max} \tag{6}$$

This implies that the production elasticity has become a time-dependent variable. Analogous relationships indicating a negative marginal product may be assumed for other production factors. Substitution of all these relationships into (3) leads to the following adjusted dynamic spatial production function:

$$\Delta y_t = (\hat{\beta} k_t + \hat{\gamma} \ell_t + \hat{\delta} e_t + \hat{\varepsilon} m_t + \hat{\xi} i_t + \hat{\eta} r_t)\,(y^{max} - \kappa y_{t-1})\, y_{t-1}/y^{max} \tag{7}$$

At first glance this may appear to be a simple non-stochastic dynamic relationship. However, it can be shown that this equation may characterize unstable and even erratic behaviour leading to periodic fluctuations.

The standard form of (7) may be rewritten as:

$$\Delta y_t = v_t(y^{max} - \kappa y_{t-1})\, y_{t-1}/y^{max} \tag{8}$$

with:

$$v_t = (\hat{\beta} k_t + \hat{\gamma} \ell_t + \hat{\delta} e_t + \hat{\varepsilon} m_t + \hat{\xi} i_t + \hat{\eta} r_t) \tag{9}$$

Equation (8) has similar properties to the well-known difference equation model analyzed by May (1974). Applications in a spatial setting may be found among others in Brouwer and Nijkamp (1985) and Dendrinos (1981). In the present context, the dynamic trajectory of the urban economy can be studied more precisely by rewriting (8) as:

$$\Delta y_t = v_t(1 - y_{t-1}/y^{max})y_{t-1} \tag{10}$$

Equation (10) is a standard equation from population dynamics and has some very interesting properties. On the basis of numerical experiments, it has been demonstrated by May (1974) that this model may exhibit a remarkable spectrum of dynamical behaviour, such as stable equilibrium points, stable cyclic oscillations, stable cycles, and chaotic regimes with a-periodic but bounded fluctuations. Two key parameters determine the stability properties of (8), viz. the initial values of y_t and the growth rate for the urban system (which depends on v_t). Simulation experiments indicate that the growth rate has a major impact on the emergence of cyclic or a-periodic fluctuations. It must be assumed that y_t is standardized (as in our model).

May has also demonstrated that a stable equilibrium may emerge if $0 \leqq v_t \leqq 2$; otherwise stable cycles or unstable fluctuations may be generated. Li and Yorke (1975) have since developed a set of sufficient conditions for the emergence of chaotic behaviour for general continuous difference equations.

Our preliminary conclusion from the above discussion is that even very simple spatial systems may be found to exhibit a rich variety of dynamic trajectories, of which balanced or cyclical growth are but two possibilities. Such heterogeneity in spatial development patterns may also be observed in the wide diversity of regional and urban trends all over the world. In our simple formulation of the urban economy, dynamic trajectories are supposedly determined by initial city size and the growth rate of the urban production system. Of course reality is certainly much more complex than this, although the fact that v_t is not a constant (but a time-dependent variable) suggests that if it is regarded as a control variable, a more stable spatial trajectory might result.

4. OUTLINE OF THE VOLUME

The above remarks and rudimentary analytical perspectives do scant justice to the many challenging problems which are associated with economic and social development over space. In this volume, a much broader collection of recent advances in the field has been assembled from leading scholars in different parts of the world.

Part A of the book focuses on the marketplace by scrutinizing the key processes of competition, specialization and technological change. Chapter 2 by Kuenne reviews the small body of literature on spatial oligopoly and suggests some lines of departure for future research in the light of current deficiencies. In the third chapter, Batten and Johansson derive a theoretical foundation for the widely-observed relative dynamics of substitution in terms of Lancaster's model of consumer behaviour. This general substitution model is remoulded to capture the dynamic processes of spatial interaction and relocation of production, thereby offering an analytical window through which a product cycle can be viewed as the result of both product substitution *and* spatial substitution proceeding in combination. In a complementary paper (Chapter 4), Aubin explores the use of viability theory to portray the notion that the evolution of a technical process of production has greater "inertia" than the evolution of the quality of the products generated by a given technology. The complementarity between Chapters 3 and 4 lies in their common concern with the nonlinear lifecycle characteristics of products or processes.

Chapter 5 continues in the vein of technical change, with the emphasis on the role of women. Chatterjee develops an agenda for research on the contemporary relationship between technology and women's employment, having explored various theoretical frameworks to analyse sex segregation in the workplace and the household economy. In Chapter 6, Hartwick reflects on Chichilnisky's suggestion that an oil importing region could benefit from an exogenous increase in the price of its imported oil. His examination of the "substitution effects" in such a model puts the economics of a booming sector on a sounder footing. The next chapter contains a rigorous derivation of the 'direct' equilibrium of an economy by Dejon, Güldner and Wenzel. By concentrating on the notion of relative demand (the quotient of purchases to supply at each price level), they are able to define a general equilibrium state without recourse to the explicit equality of supply and demand levels. Part A is completed by a study of rivalrous consonance by Kuenne. This chapter is a clear example of the suggestions made in Chapter 2.

In Part B, the focus shifts to spatial interaction analysis. Puu begins this section with an extension of his earlier work in which Hotelling's (1921) model for population growth and dispersal was given an economic rationale. By including capital accumulation as an explicit process in the model, the resulting two-input production function produces a spatially homogeneous solution which is unstable if the capital stock exceeds a certain critical value. Tony Smith follows up in Chapter 10 with a finite approach to multivariate spectral analysis. His chosen process, designated as the *circular smoothing* of the given sample, is shown to have a finite spectral representation which may be derived by simple matrix methods. In Chapter 11, Haag endeavours to improve the construction of the classical spatial interaction model by adopting a master equation approach. This type of model is also the focus of some discussion in the final chapter within Part B, where Barentsen and Nijkamp distinguish three levels on which nonlinear dynamic processes can be modelled and the resulting models classified.

The focus of Part C is urban and regional infrastructure. Mills and Carlino open the collection with an exploration of the determinants of county growth in the USA from 1970 to 1980. Considerable evidence favouring central city location over nonmetropolitan location is uncovered, emphasizing the junctional importance of the interstate highway system. In Chapter 14, Andersson and Kobayashi develop an equilibrium theory for the analysis of accessibility and density distributions. By examining nodal congestion and impacts on the resulting spatial equilibrium of population, an earlier model is given a more

general framework. The spatial equilibrium approach is again the focus of attention in Chapter 15, where Nagurney presents a general dynamic spatial price equilibrium model with gains and losses.

In Chapter 16 Lakshmanan examines the concept of infrastructure, its attributes and its role in long-term transformations of the economic system. He emphasizes the key role which network infrastructure will play in the transition to a service economy. On the other hand, Quigley and Varaiya analyze the problem of a consumer-investor who must choose a level of initial investment in insulation and other energy-saving capital to produce a flow of housing services over some time horizon. Crucial to the optimal level of investment is the expected increase in future energy prices. Rogner's paper on dynamic energy complex analysis for metropolitan planning (Chapter 18) represents a continuation of the energy theme. It emphasizes the need for a comprehensive systems approach encompassing both energy production and consumption. The book concludes with a dynamic model for the behaviour of a profit-taking developer in the housing market (Chapter 19). Kobayashi, Zhang and Yoshikawa employ the concept of infinitesimal transformations in Lie group theory to describe the taste changes of households, thereby deriving some conservation laws in the market for housing.

The above collection of writings constitutes but a small step along the path towards a better understanding of the many aspects of spatial change. Major problems which may command our attention in the coming decade concern the interdependence between technological change, knowledge creation and the formation of urban as well as internodal network infrastructure.

REFERENCES

Adelman, I., 1965, "Long Cycles - Fact or Artifact?", *American Economic Review*, June, 44:444-463.

Allen, P.M. and M. Sanglier, 1979, "A Dynamic Model of Growth in a Central Place System", *Geographical Analysis*, 11:26-272.

Alonso, W.A., 1971, "The Economics of Urban Size", *Papers of the Regional Science Association*, 26:67-83.

Andersson, Å.E., 1985, *Creativity and the Future of the Metropolis*, Prisma Press, Stockholm, (in Swedish).

Andersson, Å.E. and R. Kuenne, 1986, "Regional Economic Dynamics" in P.Nijkamp and E. Mills, (eds.), *Handbook of Regional and Urban Economics*, vol. I, North-Holland, Amsterdam, pp. 201-253.

Batten, D.F., 1982, "On the Dynamics of Industrial Evolution", *Regional Science and Urban Economics*, 12:449-462.

Batten, D.F., 1985, "The Changing Economic Structure of Metropolitan Regions", *Scandinavian Housing and Planning Research*, 2:207-223.

Batten, D.F., J. Casti and B. Johansson (eds.), 1987, *Economic Evolution and Structural Adjustment*, Springer-Verlag, Berlin.

Beaumont, J.R. and P.L. Keys, 1982, *Future Cities: Spatial Analysis of Energy Issues*, John Wiley, New York.

Biehl, D., 1980, "Determinants of Regional Disparities and the Role of Public Finance", *Public Finance*, 35:55-771.

Bluestone, B. and B.T. Harrison, 1982, *The De-industrialization of America*, Basic Books, New York.

Brouwer, F. and P. Nijkamp, 1985, "Qualitative Structure Analysis of Complex Systems", in P. Nijkamp, H. Leitner and N. Wrigley (eds.), *Measuring the Unmeasurable; Analysis of Qualitative Spatial Data*, Martinus Nijhoff, The Hague, pp. 365-384.

Dendrinos, D.S. (ed.), 1981, *Dynamic Non-linear Theory and General Urban/Regional Systems*, School of Architecture and Urban Design, Lawrence, Kansas.

Freeman, C., J. Clark and L. Soete, 1982, *Unemployment and Technical Innovation*, Frances Pinter, London.

Hirschman, A.O., 1958, Strategy of Economic Development, Yale University Press, New Haven.

Hohenberg, P.M. and L.M. Lees, 1985, *The Making of Urban Europe: 1000-1950*, Harvard University Press, Cambridge, Mass.

Hotelling, H., 1921, "A Mathematical Theory of Migration", M.A. Thesis (later version published in *Environment and Planning A*, 1978, vol. 10).

Isard, W., 1956, *Location and Space-Economy*, MIT Press, Cambridge, Mass.

Isard, W. and P. Liossatos, 1979, *Spatial Dynamics and Optimal Space-Time Development*, North-Holland, Amsterdam.

Jacobs, J., 1977, *The Death and Life of Great American Cities*, Vintage Books, New York.

Johansson, B. and P. Nijkamp, 1987, "Analysis of Episodes in Urban Event Histories", in L. van den Berg, L.S. Burns and L. Klaassen (eds.), *Spatial Cycles*, Gower, pp. 43-66.

Johansson, B. and L. Westin, 1987, "Technical Change, Location and Trade", *Papers of the Regional Science Association*, 62:13-25.

Kleinknecht, A., 1986, "Innovation Patterns in Crises and Prosperity", Ph.D. diss., Free University, Amsterdam.

Lesuis, P.K.J., F. Müller and P. Nijkamp 1980, "An Interregional Policy Model for Energy-Economic-Environmental Interactions", *Regional Science and Urban Economics*, 10, no. 3, pp. 343-370.

Li, T. and J.A. Yorke, 1975, "Period Three Implies Chaos", *American Mathematical Monthly*, 82:985-922.

Mandel, E., 1980, *Long Waves of Capitalist Development*, Cambridge University Press, Cambridge.

Markusen, A., P. Hall and A. Glasmeier, 1986, *High Tech America: the What, How, Where and Why of the Sunrise Industries*, Allen and Unwin, Boston.

May, R.M., 1974, "Biological Populations with Nonoverlapping Generations", *Science*, 186: 645-647.

Mees, A., 1975, "The Revival of Cities in Medieval Europe", *Regional Science and Urban Economics*, 5:403-425.

Mensch, G., 1979, *Statemate in Technology*, Ballinger, Cambridge.

Nelson, R.R. and S.G. Winter, 1982, *An Evolutionary Theory of Economic Change*, Belknap Press of Harvard University Press, Cambridge, Mass.

Nijkamp, P., 1986, "A Multidimensional Analysis of Regional Infrastructure and Economic Development", in Å.E. Andersson, W. Isard and T. Puu (eds.), *Regional and Industrial Development Theories, Models and Empirical Evidence*, North-Holland, Amsterdam, pp. 267-294.

Nijkamp, P., 1986, "Technological Change, Policy Response and Spatial Dynamics", in D.A. Griffith and T. Lea (eds.), *Evolving Geographical Structures*, Martinus Nijhoff, The Hague, pp. 75-99.

Norton, R.D. and J. Rees, 1979, "The Product Cycle and the Spatial Decentralization of American Manufacturing", *Regional Studies*, 13:141-151.

Olson, M., 1982, *The Rise and Decline of Nations*, Yale University Press, New Haven.

Pred, A.R., 1966, *The Spatial Dynamics of U.S. Urban-Industrial Growth, 1800-1914*, MIT Press, Cambridge.

Richardson, H.W., 1973, *The Economics of Urban Size*, D.C. Heath, Lexington.

Robson, B.T., 1973, *Urban Growth: An Approach*, Methuen, London.

Rosenberg, N.,1976, *Perspectives on Technology*, Cambridge University Press, Cambridge.

Rothwell, R., 1979, "Small and Medium Sized Manufacturing Firms and Technological Innovation", *Management Decision*, 16:362-370.
Schmookler, J., 1966, *Invention and Economic Growth*, Cambridge University Press, Cambridge.
Schumpeter, J.A., 1939, *Business Cycles*, McGraw-Hill, New York.
Sveikauskas, L., 1979, "Interurban Differences in the Innovative Nature of Production", *Journal of Urban Economics*, 6:216-227.
Wilson, A.G., 1981 *Catastrophe Theory and Bifurcation*, Croom Helm, London.

PART A

COMPETITION, SPECIALIZATION AND TECHNOLOGICAL CHANGE

Advances in Spatial Theory and Dynamics
Å.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 2

The Dynamics of Oligopolistic Location: Present Status and Future Research Directions

R.E. Kuenne

1. INTRODUCTION

The purpose of this paper is a modest one. It is to review the rather small body of literature on spatial oligopoly, judge the degree of progress it has made in yielding insights into spatially conditioned decision making in relevant economic contexts, and to suggest some lines of departure for research in the light of its present revealed deficiencies.

By "dynamic" I simply mean strategic reaction through time on the part of knowledgeable rivals that may or may not converge to a steady solution. Spatially conditioned decisions are locations and prices. Relevant contributing characteristics of the environment are number of firms, configurations of the space involved, demand function characteristics and the spatial density of demand, and firms' conjectures about their rivals' reactions.

After an extremely rich research history, which has seen pioneering applications of the calculus of variations, simulation, and catastrophe theory, among other innovations in economics, where does the theory of spatial oligopoly stand today? Why has so little been done in the last ten years? Have the various strands of investigation converged to common conclusions about the desirable path of future research? What are the paths indicated? It is these questions I should like to address in this paper, with an emphasis upon the bearing of past research upon the future.

2. THE PRESENT STATUS OF SPATIAL OLIGOPOLY ANALYSIS

In existing spatial oligopoly analysis interest focuses upon two questions: (1) the locational structure of the steady state equilibria, if they exist, and (2) the convergence or nonconvergence of the reactive process to such steady states. Most of the early work centers upon the first question, but as the inherent complexity of the real oligopoly market structure intruded into later analysis, simulation techniques had to be used. In such modelling the two problems merge. If, in a reaction sequence, when each firm is given the opportunity to adjust while its rivals remain quiescent, the solution cycles without converging, doubt is cast on the existence of a steady state that closed analysis cannot isolate. If it does converge consistently to one or a few configurations from widely different starting points, some support is given to the hypothesis that they are equilibria.

Following the lead of Gannon (1972, 1973), although generalizing his concept, I may define one body of research that originated in Hotelling's classic article (1929) as "simple spatial oligopoly". Its notable features are simplicity of demand structure and of conjectures of rival responses to initiating moves. This set of models was treated definitively in Eaton and Lipsey (1975) for the following assumptions:

1. the spatial configuration is linear, circular, or a disc, with consumers distributed uniformly or as a density function that is integrable;
2. each consumer purchases one unit of good from the firm with lowest delivered price, so that sales are proportionate to consumers in the firms' market areas;
3. all firms charge the same f.o.b. prices, with transport costs that increase with distance, and with identical, fixed marginal profit;
4. no more than one firm can occupy a location and relocation costs are zero;
5. firms locate to maximize profits subject to two alternative conjectures:
   (i) Cournot, or that rivals will not respond to initiating actions;
   (ii) game theoretic, or that rivals will respond in manners that maximize the loss of the initiating firm's market;
6. firms relocate in a dynamic quiescent sequence with the steady state identified with a Nash equilibrium.

The results of the Eaton-Lipsey analysis are too complex to be related in complete detail in this short paper, but the broad outlines are relevant for my later conclusions. For linear or circular market spaces, under Cournot expectations, the necessary and sufficient conditions for a steady state solution may be stated quite simply. Each firm will have a long side and a short side to its market segment. Then the conditions are as follows:

Condition 1: No firm's total sales are less than some firm's longside sales.
Condition 2: Peripheral firms are always paired when they occur, so their short sides are (nearly) zero in size.
Condition 3: For every unpaired firm the number of customers at its left- and right-hand boundaries are equal.
Condition 4: For every paired firm the number of customers at the short-side boundary is no less than the number at the long-side boundary.

Hotelling left the impression that spatial oligopoly competition had a tendency to cluster firms in space, or, more generally, to minimum differentiation in any dimension with spatial analogue (political programs, product brands, etc.), a hypothesis that Chamberlin (1948) sensed would not be general. Eaton and Lipsey reveal that Hotelling's pairing of firms at the center of the linear market does occur for $n = 2$ under Cournot expectations. But for $n = 3$ no equilibrium exists: Chamberlin, in asserting a contrary position, failed to grasp the significance of Condition 2. For $n = 4$ pairing occurs at the first and third quartiles, and for $n = 5$ a steady state exists with pairs at locations 1/6 and 5/6 and an unpaired firm at the center. But, for $n > 5$ an infinite number of steady states exists under simple oligopoly. If, following the suggestion of Chamberlin, we bend the line into a circle, for $n = 2$ all configurations are equilibria and for $n > 2$ an infinite number of solutions exists. By no means, therefore, is there a general tendency toward the pairing of rivals or of minimum differentiation.

As the models are increased in complexity, the indeterminacy becomes more general. In 2-space, when the market area is confined to a disc with uniform population distribution and Cournot conjectures, the solutions, if they exist, could not be obtained for $n \geq 3$, and simulation was necessary. Löschian hexagons, squares, and rectangles were not sustainable as steady state market area patterns. For $3 \leq n \leq 17$ the dynamic adjustment process failed to converge, and Eaton and Lipsey conjecture that no Nash equilibrium exists for $n > 2$ in simple spatial oligopoly. Shaked (1975) has provided a nonexistence proof for $n = 3$.

The first step to what I shall call "complex spatial oligopoly" models was taken by Lerner and Singer (1937) and, more importantly, by Smithies (1941), with the introduction of more complicated demand assumptions and the consequent introduction of price as an endogenous variable simultaneously determined with location.

Smithies's research broadened and deepened the Hotelling analysis by introducing linear demand functions, joint determination of prices and locations, and more complicated patterns of price and location conjectures, although these latter remained symmetrical for both rivals. He derives a measure of the strength of the forces pulling duopolists toward the poles of a linear market by combining the transport rate, the size of market, and a surrogate for demand elasticity into a single parameter ξ. He related location to values of this measure under three different conjecture assumptions. The relation of price to ξ was too complex to be subjected to closed analysis, although he found a general tendency for firms to absorb freight costs as they anticipated greater competitive reactions from rivals.

Gannon (1972, 1973) extended Smithies's work by introducing nonlinear demand functions and explicit marginal expectations of rivals' responses. In general, like Smithies, the burden of Gannon's work is that Hotelling's determinateness disappears with the introduction of these complications and general propositions in the absence of specified functions and parameters are not possible. Once more, in a framework of essential simplicity, the resources provided by general deductive frameworks are insufficient to yield insights.

Another predictable line of attack for spatial oligopolistic analysis in both 1-space and 2-space has been more extensive applications of game theory than the minimax conjectural variation assumption of Eaton and Lipsey permits. Stevens (1961) modified the Hotelling problem by assuming a finite number of points on the line at which location could occur, and, to make the game zero-sum, assumed that the payoff function to firms was the difference in their sales at alternative locations. Straightforward two-person, zero-sum attack upon the spatial duopoly problem concludes that a minimax solution with pairing at the center is consistent with Hotelling's results. The work was an interesting introduction of game theoretic conjectures into an essentially Hotelling framework.

More extensive applications of game theory were made by Isard and Smith (1967, 1968). In a Hotelling economy with linear demand curves, they use standard isoprofit contour analysis to derive a Cournot solution to the duopoly problem with each firm located one-third of the distance from the end-points of the linear market. Of course, both firms can benefit from a collusive agreement, and the authors consider various plausible dynamic procedures to attain such agreement. They also analyze the possibility of side payments, beginning at the joint profit maximization solution (at one of the quartiles) and negotiating over the division of the spoils. The Weber agglomeration economies problem is melded with the Hotelling case by introducing the former into the analysis. Most interestingly, the Weber agglomeration case is analyzed on its own, with various cooperative dynamic movement schemes discussed that iteratively reduce the space within which agglomeration can occur. The later article introduces the possibility of coalition formation into this agglomeration analysis.

As in all realistic oligopoly analysis, the strength of game theory ironically inheres in its disappointing results, for it succeeds in highlighting the richness of realistic processes through the methodology's general indeterminateness. It serves to reinforce the conclusions of our earlier discussions: spatial oligopoly presents dynamic adjustment potentialities that in the general case are essentially indeterminate in outcome. Hence, such problems must be accepted as *sui generis*, work is most profitably confined to specific cases via simulation, and the theoretical ambitions of the analyst should be constrained to the limited applicability of the theorems and conjectures that emerge from such analysis.

Such conclusions are not happy ones for the economic theorist. We are too much fascinated by the universal theorems of the physical scientist. The role of pure theory in illuminating applied research is a different one in the simpler universe of the physicist.

When, in our own field, it points in the direction of simulation by indicating the essentially noncorrectible oversimplicity of our methods, we simply walk away. When, so to speak, we find our theory of hydrodynamics cannot explain the flow of air around an aircraft, we do not stoop to build wind tunnels. We simply do without aircraft as not worth the condescension from respectable abstraction necessary to achieve them.

3. SUGGESTED RESEARCH DIRECTIONS

A rather compelling theme emerges from this record of research progress, in my opinion. It is that pure theory has reached the limits of its ability to illuminate spatial oligopoly using the traditional frameworks available to the oligopoly analyst. Newer methodologies must build upon the foundation of the old, but wisely. These methods must include specific assumptions about demand functions and costs, sacrificing generality of results for evaluable solutions. They must be operational, with parameters capable of estimation. And they should be flexible - more so than the rigid models of traditional oligopoly theory - with the capability of treating a broad spectrum of industry mixtures of competition and cooperation and of tailoring to comply with specific industry structures and folkways. And, lastly, perhaps more susceptible of indictment as pure personal preference, I should like to see the insights obtained when spatial oligopoly is placed within the matrix of stochastic processes, or at least when random behavior is incorporated into spatial demand.

I have some ideas about each of the three directions - simulative theorizing, cooperative rivalry, and stochastic process - but they are only investigatory beginnings or contemplations. I discuss each briefly below.

3.1 Simulative Theorizing

In a published article (Kuenne, 1977) I experimented with a simulation using n firms serving m discrete market points in space, where the firms can locate anywhere within the convex hull of the market points. Various metrics were also used to permit economic distances to approach more closely those resulting from concave transport cost functions over distance. Costs are constant for each firm but may differ among firms, linear demand functions are assumed at sinks in terms of delivered prices, which in turn are f.o.b. prices at sources plus transport costs proportionate to distance.

The resulting model is a very complicated nonlinear programming model with no constraints other than definitional identities and non-negativity constraints. The objective function - arbitrarily chosen as the maximization of joint profit - is not jointly convex in the endogenous variables and hence yields only local optima. Moreover, it is very difficult to solve, due entirely to the presence of the spatial dimension. It was solved by iterative use of two algorithms. The first assumes prices fixed and employs a multisource Weber point branch-and-bound algorithm to derive tentative source locations. The second assumes those locations to be fixed and determines optimal prices in a straightforward nonlinear programming application. Each of the two algorithms will converge to optimal solutions given the fixity of price or source location, since branch-and-bound is an exhaustive search technique and the programming problem is convex in prices. The iterative processes are stopped when convergence criteria are met.

At least, the branch-and-bound algorithm converges for small to medium size numbers of sources and sinks, and obtains very good answers for larger numbers. The price model was solved by the Sequential Unconstrained Minimization Technique (SUMT), a penalty function algorithm which is readily adapted to sequential quiescent sequences. Our

solutions converged well for 4 to 6 firms and 8 to 10 sinks, approaching the same solutions for widely varying initial assumptions.

The model was used to exploit Smithies's insight into the importance of transport rate, size of market, and demand elasticity in spatial oligopoly, or the $\underline{s}$ measure discussed in Section 2. Also the model was used to measure changes in locational patterns and prices when joint profit maximization is compared with purely competitive pricing and Cournot-myopic expectations of rivals' responses.

One difficult problem occurs in extending this framework to permit multi-objective decision making in a spatial context by including constraints *in the spatial dimension.* Those objectives that impact only prices directly are easily handled, but combinatorial programming algorithms like branch-and-bound find it difficult to deal with constraints. Hence, constraints on location, which exist in the realistic location decisions of firms, are difficult to include. Research into means of incorporating them into branch-and-bound algorithms, or into developing alternative heuristic algorithms as substitutes, would be extremely valuable.

In a deeper sense, we also need a better framework for efficient parameter manipulation of such simulations to derive insights of maximum generalizability. Are there general methods for quickly discerning the sensitive parameters in such models? Can the notion of Smithies's artificial $\underline{s}$ parameter that economically combines the influences of several parameters be generalized? Are there general methods for placing upper and lower bounds on parameters? Must each sensitivity analysis be an ad hoc procedure without general principles of structure, formulation, or interpretation, or can general frameworks be found to generalize such analysis?

We truly require an "econometrics of simulation" that specifies general guidelines for structuring and codifying models and post-optimality analysis, if simulative theorizing is to become a tool of theoretical research.

3.2 Rivalrous Consonance

In recognition of the varying mixtures of rivalry and cooperation that characterize market structures, ranging from joint profit maximization through Cournot disregard of rivals' profits into active warfare, the simulative joint-profit maximization objective can be decentralized to permit each firm to maximize its own profit plus the weighted sum of rivals' profits. The weights reflect the firm's conception of the industry power structure in the broadest sense from its own viewpoint. Other goals can be incorporated as constraints if the algorithmic difficulties discussed above were solved.

I have advocated this framework in nonspatial oligopolistic studies for a variety of reasons, ranging from its ability to model mature oligopolistic industry behavior more realistically to the possibility of developing a theory of general oligopolistic equilibrium (see Chapter 8 in this volume). It would be most interesting to see what its application in the spatial dimension would imply. How is tacit cooperation or bounded competition reflected in location patterns? In what manners do they change as rivalry increases? What is the locational equivalent of active price war, set off when firms value rivals profits at negative values? At the present time I am most interested in making this extension of the rivalrous consonance framework, with and without the multi-objective complication. For example, as a start, it would be interesting to model the Eaton-Lipsey cases on the disc for n between 3 and 17. One could approximate the assumption of a uniform continuous distribution of demand by increasingly dense evenly distributed discrete demand points, and use the multisource Weber point and optimal pricing algorithms in tandem for solutions. This might confirm or refute the conjecture of nonequilibrium.

3.3 Stochastic Process Theory

Stochastic process theory has been extensively used in spatial economics only in the adoption of diffusion theory to explain the spread and patterning of innovations and similar phenomena over space. My surmise is that it could be usefully employed in explaining the nature of certain types of contingency demand over space, the evolution of market areas, the sequence of rival entries, and the resulting spatial patterning of locations and prices. No doubt, as is true of stochastic processes with any degree of complexity, simulation would have to be resorted to once more. However,the discernment of steady state patterns and their changes under different parametric regimes could be quite rewarding.

My only excursion into such areas has been an investigation of an elementary Poisson demand process in space and its implications for the market area of a single firm (Kuenne, 1984). If the occurrence of demands in a time period over space were Poisson (which I assumed) and the amount of demand with such an occurrence followed a geometric distribution (which I did not assume), then demand over space is a "stuttering Poisson" process with a negative binomial probability function. It is well known that for a sufficiently large number of occurrences the normal distribution approximates this function, and it would be relatively simple to study the nature of oligopolistic decision making under such regimes, when joint profit maximization, Cournot conjectures, or, more broadly, rivalrous consonance, is assumed.

4. CONCLUSION

Spatial oligopoly theory developed along several paths from Hotelling's original formulation of a linear economy with perfectly inelastic demand. These included differently shaped spaces, with and without interiors, elastic demand functions, game theoretical rivalry, the endogenization of prices, and non-Cournot conjectural variation. Some of the efforts were definitive, notably that of Eaton and Lipsey's extension and generalization of the Hotelling problem. All have explicitly or implicitly reached the conclusion that general deductive analysis with nonspecified models could go no further profitably.

The field of spatial oligopoly analysis, as a consequence, has been dormant for the last decade or so. There is a clear need to stimulate analysis in this most relevant market structure as it relates to space. The need to investigate simulative alternatives with specified parameters and functions seems clear. Among other advantages it will permit the investigation of differentiated oligopoly, blends or rivalry and cooperation, multi-objective decision making and larger numbers of rivals. Most importantly, it offers the only presently perceived route out of the impasse that current methods have engendered.

REFERENCES

Chamberlin, E.H., 1948, *The Theory of Monopolistic Competition*, 6th edition, Harvard University Press, Cambridge, M.A.

Eaton, B.C.and R.G. Lipsey, 1975, The Principle of Minimum Differentiation Reconsidered: Some New Developments in the Theory of Spatial Competition, *Review of Economic Studies*, 45:27-29.

Gannon, C.A., 1972, "Consumer Demand, Conjectural Interdependence, and Location Equilibria in Simple Spatial Duopoly", *Papers of the Regional Science Association*, 28: 83-107.

Gannon, C.A., 1973, "Central Concentration in Simple Spatial Duopoly: Some Behavioral and Functional Conditions", *Journal of Regional Science*, 13:357-375.

Hotelling, H.,1929, Stability in Competition, *Economic Journal*, 39:41-57.

Isard, W. and T.E. Smith, 1967, "Location Games: With Applications to Classic Location Problems", *Papers of the Regional Science Association*, 19:45-80.

Isard, W. and T.E. Smith, 1968, "Coalition Location Games: Paper 3", *Papers of the Regional Science Association*, 20:95-107.

Kuenne, R.E., 1977,"Spatial Oligopoly: Price-location Interdependence and Social Cost in a Discrete Market Space", *Regional Science and Urban Economics*, 7:339-358

Kuenne, R.E.,1984, "Economic Decision Making in a Poisson Demand Space", in Å.E. Andersson, W.Isard, and T. Puu (eds.), *Regional and Industrial Development Theories, Models and Empirical Evidence*, pp. 331-346, North-Holland, Amsterdam.

Lerner, A.P. and H.W. Singer, 1937, "Some Notes on Duopoly and Spatial Competition", *Journal of Political Economy*, 45:145-160.

Shaked, A., 1975, "Non-existence of Equilibria for the Two-dimensional Three-firm Location Problem", *Review of Economic Studies*, 45:51-55.

Smithies, A., 1941, "Optimal Location in Spatial Competition", *Journal of Political Economy*, 49:423-439.

Stevens, B.H., 1961, "An Application of Game Theory to a Problem in Location Strategy," *Papers of the Regional Science Association*, 7:143-157.

Advances in Spatial Theory and Dynamics
Å.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 3

Dynamics of Product Substitution

D.F. Batten and B. Johansson

1. INTRODUCTION

Industrial sectors in most nations of the world are undergoing gradual but major technological changes. Although most consumers still have relatively few *basic* needs - such as food, clothing, shelter, transportation, education, employment and the like - there are myriads of changes occurring at the intermediate stages of production as well as in the individual choice processes of households. Regardless of whether we consider intermediate or final users, advancing sophistication and technological evolution consist mainly of substituting new means of consumer satisfaction for old ones. The basic needs do not undergo radical change, but the detailed ways and means of satisfying them may alter markedly. Under these circumstances, the notion of competitive substitution as a model of technological change will apply.

There is ample evidence to suggest that, under freely competitive conditions, the dynamic processes of market penetration and product substitution follow S-shaped curves. The trajectory of the product cycle generally portrays sales growth as exponential during a product's formative years, with an inevitable slowdown as the product matures and new competitors enter the market. Quite often, the logistic curve can provide a useful framework for modelling the substitution process in relative terms (see, for example, Fisher and Pry 1971; Peterka 1977; Batten and Johansson 1985a). This type of observation has been used as a constraint on the model framework developed in the subsequent analysis, which comprises the following aspects: (i) products, characteristics and substitution, (ii) dynamics of supply and demand, and (iii) deliveries over geographical space.

1.1 Products, Characteristics, and Competitive Substitution

Our initial objective is to derive a theoretical foundation for the substitution process (under very general conditions) in terms of Lancaster's model of consumer behaviour. In this respect, the preference functions of consumers and producers are assumed to rank products (or groups of products) via the characteristics they possess. The specific form of each customer group's preference function in this product-characteristic space (and thus his sensitivity to price variations) determines the feasible substitution possibilities.

By introducing a space of characteristics (or attributes) we obtain an invariant structure in which the dynamic competition between products take place. The dynamics in the model has the form of adjustments to changing conditions such as development of relative prices, production costs, attributes of products, and preferences. These adjustments generate temporal patterns of a sigmoid nature, and are initiated by instabilities and discontinuities in supply and demand behaviour. Such unstable situations have a catalytic influence on the entry into a market by entirely new products. Partial or complete penetration result depending on the composition of different customer groups. In this respect, the formulations extend and enrich some earlier work on industrial evolution (Batten, 1982) by integrating studies of product competition (Johansson, 1978) into a structure of differential equations describing adjustments to changing prices and product characteristics.

1.2 Product Cycles, Interregional Trade, and Relocation in Space

The second purpose of the paper is to refine the general substitution model in order to capture the dynamic processes of spatial interaction and relocation of production. In broad terms, a product cycle can be viewed as the result of both product substitution and spatial substitution proceeding in combination. The evolution of spatial cycles of production in terms of customer choice, regional specialization and changing regional market shares can therefore be explored using a similar system of nonlinear differential equations.

A fundamental feature of such change is the gradual spatial relocation of production centres between regions and an associated gradual or abrupt transformation of trade flows. Relocation and adjustment of delivery patterns can be seen as a response to the penetration of new competitors into the domestic and export markets and a resulting desire to find profitable trade links and cost-reducing locations. In this respect, the framework outlines a "duality" between the behaviour of customers ranking collections of product characteristics and the behaviour of firms facing decisions about product mix, trade and location. Each can be seen to influence the competitive substitution of more for less favoured products over space and time. The paper ends by providing a few empirical examples which illustrate some dynamic processes derived from the analyses.

2. PRODUCTS, CHARACTERISTICS AND SUBSTITUTION

In Lancaster's model of consumer behaviour, products (or goods) possess objectively measurable characteristics (or attributes). Customers buying a specific product generally comprise both households, firms and other organisations. In principle all these various customers use products, singly or in combination, as inputs to a production or consumption process. Preference orderings are assumed to rank collections of characteristics and only to rank products or collections of products indirectly via the characteristics they possess (c.f. Lancaster, 1971).

2.1 Preferences of Customer Groups

We consider a product group to consist of products which (i) have certain characteristics in common, and (ii) may be perceived as satisfying a similar purpose or need. In the present study we do not deliberate upon the interdependence between different product

groups. Associated with such a group is a market in which any product may be displaced by a new product with a preferred bundle of characteristics or as a consequence of relative price changes. In this market we identify $n \geq 1$ products $i \in I$, where I is the index set of the group. Let

$$x = (x_1, ..., x_n) \in X = \{x \in R^n : x \geqq 0\} \tag{1}$$

be a vector of products, where x_i denotes a quantity of product i. Let

$$z = (z_1, ..., z_m) \in Z = \{z \in R^m : z \geqq 0\} \tag{2}$$

be a vector of characteristics, where z_k denotes a quantity of characteristic k. Moreover, let C be an attribute mapping from X to Z which relates every vector of products to a vector of characteristics so that

$$z = C(x) \tag{3}$$

Finally, let $u(z)$ signify a real-valued preference function which describes a customer's evaluation of z-vectors, and assume that each customer chooses a situation which maximizes $u(z)$. The preference structure may vary in character for different types of customers and may also change over time due to changes in the production techniques of producers using products $x \in X$ as inputs. As a point of reference we adopt the following orthodox assumption

> u is continuous, strictly quasi-concave and differentiable with all first order derivatives positive (all characteristics positively desired). (4)

For each market (product group) we recognize the set G of customer groups. Each group $g \in G$ is identified by its preference function u^g and budget m^g which at each point in time is allocated to purchase products from group I.

2.2 Customer Choice and Demand Functions

The set-valued mapping K describes the choice set of customer group g

$$K(p,m^g) = \{x \in X : px \leq m^g\} \tag{5}$$

where $p = (p_1, ..., p_n)$ and p_i is the price of product i; px denotes the inner product of p and x. The amount of different characteristics (attributes) associated with one unit of product i is described by a vector $b^i = (b^i_1, ..., b^i_m) \geq 0$ with at least one $b^i_j > 0$.

If every customer can combine any of the products to obtain a desired amount of characteristics, each vector x is assumed to be transformed to a characteristic vector $z = C(x)$ as follows

$$C(x) = Bx \tag{6}$$

where B is an irreducible matrix with b^i as its i'th column vector. If products cannot be combined, the following transformation applies

$$z \in \{ C(x_i) , i = 1, ...,n \} \quad (7)$$

where $C(x_i) = b^i x_i$ (compare Lancaster, 1982). With the specification in (6) the demand of customer group g can be expressed as

$$F^g (p,m^g) = \{ x \in K (p,m^g) : u^g (Bx) = \max u^g (Bx) \} \quad (8)$$

Application of (7) yields the demand

$$\hat{F}^g (p,m^g) = \{ x \in K (p,m^g) : u^g (C(x_i)) = \max u^g (C(x_i)) \} \quad (9)$$

Both F^g and $\hat{F}^g$ may be set-valued. An element of F^g is a vector with components

$$x_i^g = v_i^g (p)\, m^g/p_i \quad (10)$$

where for a given budget v_i^g is a function describing the share of m^g which is spent by group g on product i. An element of $\hat{F}^g$ has the form $(0, ..., x_i^g, ..., 0)$. Formula (10) also applies in this case.

Observe that m^g/p_i is the maximum amount of product i that group g is able to purchase. Hence, $0 \leq v_i^g(p) \leq 1$. Moreover, $K (p,m^g)$ is a bounded set (convex polyhedron). Then one may conclude from (8) and (10)

Lemma 1 Let $v^g (p) = (v_1^g (p), ..., v_n^g(p))$. If $F^g (p, m^g)$ is vector-valued, $v^g (p)$ is unique. Assume that (4) and (6) applies. Then F^g is upper semi-continuous with convex images (see Berge, 1959, pp 122-123). The same property is naturally valid also for

$$F(p,m) = \sum_{g \in G} F^g (p,m^g) \quad (11)$$

Consider a customer group for which (7) applies, which means that goods cannot be combined. In this case $v_i^g (p)$ is either zero or one. As long as $v_i^g (p) = 1$, x_i^g varies continuously with p_i. However, when p_i is gradually increased, and all other prices are kept constant, the budget share, $v_i^g (p)$, will drop abruptly and discontinuously to zero, while $v_j^g (p)$ simultaneously jumps from zero to unity for some other product $j \neq i$. At the point of discontinuity, $p = p^*$, $v_i^g(p^*) = v_j^g (p^*) = 1$.

Let $m = \Sigma_g m^g$ and formulate the following overall budget share function v_i such that

$$v_i (p) = \sum_{g \in G} v_i^g (p)\, m^g/m \quad (12)$$

Then we observe

> Remark 1 Assume that (7) and (9) apply. Then v_i exhibits the same type of discontinuities as v_i^g as long as the market consists of a finite number of customer groups. As the number of such groups with differentiated preference functions u^g increases, the relative size of the discontinuous jumps in v_i will shrink.

Remark 1 pertains to customers who buy only one product to achieve a desired bundle of characteristics during a given time interval. In the sequel the analysis is focussed primarily on cases where a customer can combine two or more products to obtain a desired constellation of characteristics. An example is a firm using several different types of equipment, such as machines and trucks. Another example is a household using one car and a caravan instead of a combi-van as a means of producing part of its transportation and accomodation services.

2.3 Excess Demand and Actual Sales

For each product i and at each point in time we identify the set G_i of customer groups with such production and consumption activities (and knowledge about such activities) that they are able to contemplate using product i as an input to their respective activities. Let $g \in G_i$ and define group g's potential demand as $D_i^g(p) = m^g/p_i$. The aggregate potential demand can then be calculated as

$$D_i(p) = \sum_{g \in G_i} D_i^g(p) \tag{13}$$

The notional demand is obtained from (12) and (13) as $v_i(p)\, D_i(p) = v_i(p)\, m/p_i$. Let x_j be the supply of product j. Then we may express the actual sales as $\beta_j x_j$, where

$$\beta_j = \min \{ 1,\ v_j(p)m / p_j x_j \}$$

Note that $m - \Sigma_j \beta_j p_j x_j$ may also be positive when the supply of some products is greater than what is demanded. Consequently, we introduce the following excess demand expression for product i:

$$E_i = D_i(p) - \Sigma_j \beta_{ij} x_j \tag{14}$$

where $\beta_{ij} = v_j(p)m / p_i x_j$ and $\beta_{ii} = 1$. In order to use the E_i-function in our dynamic specification of market development and product substitution, we need to establish certain properties of this function. Lemma 2 prescribes conditions which guarantee stronger continuity properties than those presented in Lemma 1.

> Lemma 2 Consider the conditions given by (4)-(6) and (8). Let the attribute vectors b^i of B in (6) be linearly independent for all $i \in I$. Then the functions v_i^g will be continuous and with unique values for each p. A proof is outlined in the appendix.

From Lemma 2 it follows that $v_i(p)$ varies continuously with p as long as all b^i-vectors are linearly independent which means that every product $i \in I$ is uniquely differentiated from the remaining products in the product group. This also implies that for fixed x_j, β_{ij} is a continuous function of prices. Naturally, the same is true for E_i.

In order to evaluate E_i we rewrite (14) as

$$E_i = m/p_i - \sum_{j \neq i} v_j(p)\, m/p_i - x_i \tag{15}$$

which may also be written as

$$E_i = m/p_i - (1 - v_i(p))m/p_i - x_i \tag{16}$$

First we use (16) to find that

$$\partial E_i/\partial p_i = (\partial v_i/\partial p_i)\, m/p_i - v_i(p)m/(p_i)^2 < 0 \tag{17}$$

The condition about the sign follows from $\partial v_i/\partial p_i \leq 0$. Our second task is to evaluate $\partial E_i/\partial p_j$ from (16), which comes out as

$$\partial E_i/\partial p_j = (\partial v_i/\partial p_j)\, m/p_i \geq 0 \tag{18}$$

The expression in (18) can be equal to zero, since for certain prices the optimization problem in (8) may yield "corner solutions". A product k may for example at some price vector p satisfy $v_k^g(p) = 1$ and $v_k^h(p) = 0$ for all $h \neq g$. This means that one customer group g is spending all its budget (for product group I) on product k, while all other groups do not buy k. It is a "sticky" corner solution, in so much as small price variations around p will not necessarily alter this situation. However, if there is at least one g such that $0 < v_i^g(p) < 1$ and $0 < v_j^g(p) < 1$, then $\partial E_i/\partial p_j > 0$. This follows from the assumption about u^g in (4). We may thus summarize the results in this subsection as follows:

Proposition 1 Given the assumptions in Lemma 2 we may conclude that

(i) $\Delta E_i \Delta p_i < 0$
(ii) $\Delta E_i \Delta p_j \geq 0$
(iii) $\Delta E_i \Delta p_j > 0$ if $0 < v_i^g < 1$ and $0 < v_j^g < 1$ for at least one g.

3. DYNAMICS OF SUPPLY AND DEMAND

In Section 2 we described how the excess demand variable E_i is affected by price alterations. Over time E_i is also influenced by the supply behaviour of firms. This is evident from formula (15). We will assume that the state $E_i < 0$ influences x_i to decrease and the state where $E_i > 0$ changes x_i in the opposite direction, provided the price p_i yields a sufficiently high profit. We may summarize this assumption as follows:

$$\dot{x}_i / x_i = \begin{cases} \alpha_i E_i & \text{if } x_i \leq \bar{x}_i \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

where $\bar{x}_i$ denotes the capacity to supply product i at each time. The coefficient $\alpha_i \geq 0$ may be interpreted as a response coefficient which is sensitive to the profit levels associated with deliveries to different regional markets (compare Mansfield, 1961). From (19) it follows that $\dot{x}_i = \dot{\bar{x}}_i$ as long as $\bar{x}_i = x_i$ and $\dot{x}_i > 0$.

3.1 Sigmoid Development Paths

Assume that all products are uniquely differentiated so that Lemma 2 applies. From (16) we know that E_i may be written as

$$E_i = (m/p_i)\, v_i(p) - x_i \tag{20}$$

from which we obtain the following form of (19):

$$\dot{x}_i = \alpha_i [(m/p_i)\, v_i(p)\, x_i - x_i^2] \tag{21}$$

which is immediately recognized as a growth path of the logistic type belonging to the wider family of sigmoid distributions. Writing $A_i(p) = (m/p_i)\, v_i(p)$, it becomes clear that $x_i = \alpha_i x_i [A_i(p) - x_i]$ will reach a stationary state at the point $x_i = A_i(p)$. Moreover, gradual reductions of p_i relative to other prices p_j mean that $A_i(p)$ will expand over time. Similarly, x_i will decline if $A_i(p) < x_i$.

Consider a gradual stepwise expansion of $A_i(p)$. Such an expansion will have a definite limit if we introduce the following assumption. First, let the unit profit be $\pi_i = p_i - c_i(\bar{x}_i)$, where $c_i(\bar{x}_i)$ is the unit cost. Assume that $\partial c_i / \partial \bar{x}_i < 0$ as $\bar{x}_i < \bar{\bar{x}}$ and that $c_i(\bar{x}_i) \geq \bar{c}_i = c_i^{min}$. Assume next that there is a profit level $\bar{\pi}$ which serves as a minimum mark up. If the price p_i falls below $c_i(\bar{x}_i) + \bar{\pi}$, no capacity investments are carried out. In this way we obtain an upper bound for $A_i(p)$.

In accordance with the above we assume that the following price-setting rule applies for a product in a phase of capacity expansion:

$$p_i(\bar{x}_i) = c_i(\bar{x}_i) + \bar{\pi} \tag{22}$$

During a phase of decline there are no new investment costs to cover. Hence, we can introduce the price setting condition $p_i \geq c_i(\bar{x}_i)$ for phases of decline. To reflect this asymmetry between expansion and decline, capacity changes may be associated with the dynamics of output in the following way:

$$\dot{\bar{x}}_i = \delta_i (\bar{x}_i, x_i) \dot{x}_i$$

$$\delta_i = \begin{cases} 0 \text{ if } x_i > \bar{x}_i > \gamma_i \bar{x}_i \\ 1 \text{ otherwise} \end{cases} \tag{23}$$

where $0 < \gamma_i < 1$, and $\gamma_i = \gamma_i (\bar{x}_i, x_i)$.

Treating α_i as a constant, we may consider the ratio $F_i = p_i x_i / (m - p_i x_i)$ to derive

$$\dot{F}_i = m (\dot{p}_i x_i + p_i \dot{x}_i) / (m - p_i x_i)^2$$

From this we obtain

$$\dot{F}_i / F_i = m (\dot{p}_i x_i + p_i \dot{x}_i) / p_i x_i (m - p_i x) \tag{24}$$

A logistic development path obtains if the right hand side of (24) assumes a constant value. In this case, Verhulst-Pearl's equality $\dot{F}_i / F_i = h_i$ is satisfied, where h_i is a constant which is positive as F_i grows. When examining this case we observe that (24) describes a simultaneous adjustment of price and quantity. Therefore we need an assumption about the motion of p_i. Let us assume that the price is adjusted in such a way that (22) is satisfied. This corresponds to a growth process. In addition we assume that $c_i(\bar{x}_i)$ is scale-dependent, which is described by the following specification:

$$c_i(\bar{x}_i) = c_i^0 \exp \{ - k_i \bar{x}_i \} + \bar{c}_i \tag{25}$$

where $c_i^0 > 0$, $k_i > 0$ and $\bar{c}_i$ is a lower bound. The price-setting rule in (22) together with (25) yields $p_i = G_i(\bar{x}_i) + \bar{\bar{c}}$, where $G_i(\bar{x}_i) = c_i(\bar{x}_i) - \bar{c}_i$ and $\bar{\bar{c}} = \bar{c}_i + \bar{\pi}$ Hence,

$$\dot{p}_i = -k_i G_i (\bar{x}_i) \dot{\bar{x}}_i \tag{26}$$

Let us consider a development path along which $\dot{x} = \dot{\bar{x}}_i > 0$, A prerequisite for this process is that p_i remains low enough (relative to other prices) to ensure that $A_i(p) - x_i$ remains positive. Let $\bar{G}_i (\bar{x}_i) = G_i (\bar{x}_i) + \bar{\bar{c}}$ and apply the price adjustment process in (26) to formula (24). Then we obtain (for $\dot{F}_i / F_i = h$)

$$h \bar{G}_i (\bar{x}_i) x_i (m - \bar{G}_i (\bar{x}_i) x_i) = -m k_i G_i (\bar{x}_i) x_i + m \bar{G}_i (\bar{x}_i) \dot{x}_i$$

Dividing both sides by x_i, we get

$$\dot{x}_i / x_i = h \bar{G}_i (\bar{x}_i) [m - \bar{G}_i (\bar{x}_i) x_i] / m [\bar{G}_i (\bar{x}_i) - k_i G_i (\bar{x}_i)] \tag{27}$$

The condition $\dot{F}_i/F_i = h$ means that market shares develop along a logistic path. Formula (27) demonstrates that this market-share behaviour also implies that the produced quantities follow a sigmoid growth path which terminates ($\dot{x}_i = 0$) as $\bar{G}_i(\bar{x}_i)\, x_i = m \geq \bar{\bar{c}}_i\, x_i$. This corresponds to complete market penetration by the (new) product i.

3.2 Market Shares

During a growth phase, (26) portrays a gradual price reduction such that profits are at each stage just sufficient to cover the associated investment costs. When production is falling, (26) and (27) taken together reveal that the price will not rise as long as $x_i > \gamma_i \bar{x}_i$.

Product i's market share $f_i = p_i\, x_i / m$ develops in the following way:

$$\dot{f}_i = (1/m)\,(\bar{G}_i(\bar{x}_i)\,\dot{x}_i - x_i\, k_i\, G_i(\bar{x}_i)\,\dot{x}_i) \tag{28}$$

This means that f_i grows when $\dot{x}_i\,[\,G_i(\bar{x}_i)\,(1 - k_i\, x_i) + \bar{\bar{c}}_i\,] > 0$; requiring that $x_i > 0$, which depends on the prices of all products. Moreover, the market share will grow only if

$$k_i < H_i(x_i) = \bar{\bar{c}}_i / x_i\, G_i(\bar{x}_i) + 1/x_i \tag{29}$$

According to (26), $k_i\, G_i(\bar{x}_i) < 1$ implies that $x_i\, G_i(\bar{x}_i)$ will grow when x_i grows. Hence, $H_i(x_i)$ will gradually decline as x_i expands. Thus k_i may exceed $H_i(x_i)$ for large values of x_i whereupon $\dot{f}_i < 0$ as $\dot{x}_i > 0$. However, for values of E_i close to zero we may write $f_i \approx A_i(p)\, p_i / m = v_i(p)$ which means that any reduction in f_i will induce a reduction of E_i which, in turn, retards the rate of increase in x_i so that $\dot{x}_i$ approaches zero. In this way, a previously expanding market share may cease either because $f_i = 1$ or because $k_i \geq H_i(x_i)$. The latter case implies an instability phenomenon. A price-reduction policy from an established product, or the introduction of a new, slightly superior product j, will temporarily induce the condition $E_i = A_i(p) - x_i < 0$. This induces a continuing fall in output which, after some delay implied by (23) and (26), forces p_i to increase. Hence, whenever E_i is almost zero, any small alterations in relative prices may cause a growth phase to be transformed into a phase of decline. From (28) it is evident that each such path will assume the shape of a sigmoid distribution.

3.3 Introduction of New Products and Catastrophic Shifts in the Dynamics

Let us consider a specific customer group. Moreoever, let there be m distinct characteristics which are positively valued by these customers. Supppose that there are n different products which are uniquely differentiated in the sense that their attribute vectors $b^1, \ldots, b^n$ are linearly independent. Let us then introduce a new product n+1 such that b^{n+1} can be expressed as a linear combination of the initial n attribute vectors. We are especially interested in the case where n+1 is intermediate in the sense that

$$b^{n+1} = \sum_{i=1}^{n} \alpha_i b^i$$

as $\alpha_i \geq 0$ for all i and $\alpha_j > 0$ for at least one j. Then product n+1 may attract all customers of those products j such that $\alpha_j > 0$. A case of this kind is illustrated in Figure 1. This case has two attributes and three products, represented by the attribute vectors b^1, b^2 and b^3. The figure describes a price constellation p^* such that $v_1^g(p^*)\, b^1/p_1^* + v_2^g(p^*)\, b^2/p_2^* = b^3/p_3^*$. Obviously, any reduction $\varepsilon > 0$ such that $p_3 = p_3^* - \varepsilon$ will cause v_1^g and v_2^g to drop to zero. Hence, if the prices of products 1 and 2 are not reduced accordingly, these products will lose customer group g.

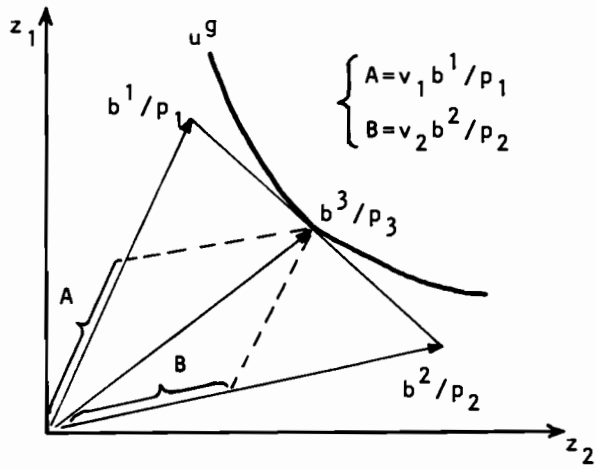


Figure 1 Introduction of an intermediate product

Proposition 2 Let the conditions in (4)-(6) and (8) be given and let the attribute vector $b^1 = \alpha_2 b^2 + ... + \alpha_k b^k$, where all $b^i \geq 0$, $b^i \neq 0$, and all $\alpha_i > 0$. Assume that b^1 contains all attributes of product group I which are relevant for customer group g. Then there exist prices p_1^* and $p_{(1)}^* = (p_2^*,...,p_k^*)$ such that (a) $E_1(p_1^* + \varepsilon, p_{(1)}^*) \leq 0$, (b) $E_1(p_1^*, p_{(1)}^*)$ is undecided, and (c) $E_1(p_1^* - \varepsilon, p_{(1)}^*) = m^g/(p_1^* - \varepsilon) - x_1 > 0$, where $\varepsilon > 0$ is arbitrarily small, and $x_1 < m^g/p_1^*$.

A proof of the proposition is given in the appendix. The proposition complements the statement in Lemma 2. In the appendix the following remark is also assessed:

> Remark 2 Let all conditions of Proposition 2 be given, except that b^1 now does not necessarily contain all attributes (in the product group) which are relevant for customer group g. The assumptions in the proposition guarantee that b^1 contains all attributes that can be obtained from a combination of b^2, ..., b^k. Suppose that $v_i(p_1^* + \varepsilon, p_{(1)}^*) > 0$ for all i + 2, ..., k. Then

(a) $E_1 (p_1^* + \varepsilon, p_{(1)}^*) = 0$

(b) $E_1 (p_1^* p_{(1)}^*)$ is undecided

(c) $E_1 (p_1^* - \varepsilon, p_{(1)}^*) = 0$

Proposition 2 and Remark 2 have strong implications for the relation between the number of products and attributes. One such implication is described in the following corollary, which is self-evident given Proposition 2.

> Corollary 1 Let the conditions of Proposition 2 apply for a product group I with n products such that $b^i \geq 0$, $b^i \neq 0$ for all $i \in I$, and $\Sigma\, b^i > 0$. Let m denote the number of attributes. If $n \geq m$, one can always construct a new intermediate product of the same kind as product 1 in Proposition 2.

Proof. The conditions in Corollary 1 imply that the attribute vectors $b^1, ..., b^n$ are linearly dependent. Hence, there exists a combination $\alpha_1 b^1 + ... + \alpha_n b^n = 0$ for $(\alpha_1, ..., \alpha_n) \neq 0$. Collect all negative terms on the RHS and all positive terms on the LHS, so that RHS = LHS. This is an appropriate vector value for the new product.

A similar type of discontinuous behaviour in the E_i-function was described for products which cannot be combined in connection with Remark 1. The background for this property is illustrated in Figure 2. Here we should note that all three vectors b^1/p_1, b^2/p_2 and b^3/p_3 satisfy the same value of customer group g's preference function. Hence, at this price constellation the F^g-function in (9) is discontinuous in such a way that if any price is altered at least one product will lose all its sales to customer group g. In Johansson (1978) this situation is analysed for the case with several distinct customer groups.

3.4 The Effect of having Distinct Customer Groups

One important effect of having several different customer groups is that discontinuous changes in demand will usually affect only one customer group at a time, (i.e., several groups in a predictable sequence, given the relative price changes). A second, interrelated aspect is that different customer groups may not perceive the same subsets of attributes to be relevant. Thus a new and superior product may not necessarily capture the whole market but may only attract the sales of certain customer groups. Formally, this means that $D_i^g(p)$ in (13) may be zero for all p-values while it is positive for $D_j^g(p)$.

To reflect the division of customer groups we may define the relevant market budget of product i as

$$m_i = \sum_{g \in G_i} m^g$$

where G_i is the set of customer groups which are potential buyers of product i. Some or all of these customers may be industrial firms. Their short term preferences depend on the production technique they apply. In the longer term information about alternative techniques diffuses to all firms and they may invest in completely new production systems. As a consequence their preferences will change. In our present setting, this can

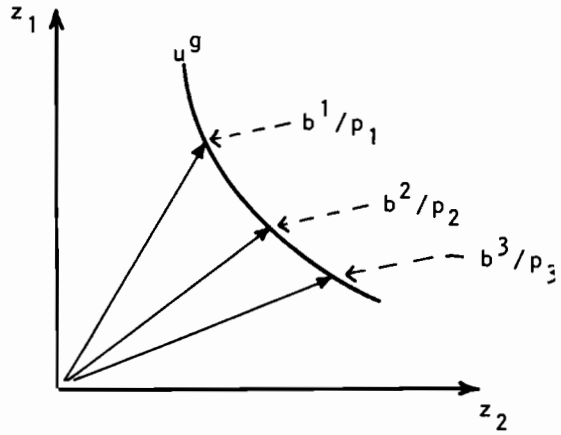


Figure 2 Point of discontinuity for products which cannot be combined

be represented as a switch to another customer group. If this type of transition is a gradual process, it may be portrayed as a smooth increase in the overall customer budget of product i so that $m_i = \lambda_i m_i (m - m_i)$. In addition the size of the market itself may grow in so much as the budget m is expanded. In this context, the division of market budgets between different product groups is affected by the development of the aggregate price levels of the different groups. We shall not discuss this inter-group competition here, but instead consider the question of competitive substitution effects over geographical space.

4. SPATIAL DELIVERIES

In the preceding section, our interest is concentrated on the competition between different products. The supply behaviour and price-setting rules do not reflect any explicit competition between different firms producing the same product. Such an approach is consistent with the view that the output of one firm is always differentiated to some extent from that of all other firms.

When we introduce different regions or locations, additional features are then essential and therefore we may allow firms in different locations to produce goods which are identified as identical. In this way, spatial aspects such as delivery costs are emphasized.

4.1 Origin-Destination Relationships

Consider a spatial economy which is identified by a set of locations r,s = 1, 2. ..., and a displacement cost coefficient c_i^{rs} for each product i and origin-destination pair r,s. Each location represents a separate market with its own price formation process - for example different regions or countries. A delivery from r to s is denoted by x_i^{rs}, and the

production (output) in r by x_i^r. For region s, we introduce the following excess demand expression

$$E_i^s = A_i^s(p^s) - \sum_r x_i^{rs} \tag{30}$$

where $p^s = (p_1^s, \ldots, p_n^s)$ is the price vector of product group I in region s; $A_i^s(p^s) = (m^s/p_i^s)\, v_i^s(p^s)$, where m^s and $v_i^s(p^s)$ denote the budget in region s and product i's budget share in region s, respectively (compare formula (15)). In accordance with our earlier analysis the deliveries from r to s are assumed to increase only if $E_i^s > 0$. However, with several regions involved the delivering region r generally faces a whole set of customer regions. We assume that each supply region has a higher propensity to deliver on links which offer high profits. More specifically, we assume that deliveries are gradually redistributed in response to the profit differentials of alternative links (r,s), (r,k), etc. The profit of a link (r,s) is defined as

$$\pi_i^{rs} = p_i^s - c_i^{rs} - c_i^r \tag{31}$$

where c_i^r denotes the production cost per unit output in region r, c_i^{rs} the delivery costs on the link r,s, and p_i^s the market price in region s.

The customer model introduced in section 2.2 necessitates one price for each product in a market. In order to retain the earlier assumption about price-setting suppliers, the market price in region s is assumed to reflect the cost structure of all suppliers in the following way (compare Batten and Johansson, 1985b):

$$p_i^s = \sum_r p_i^{rs} x_i^{rs} / \sum_r x_i^{rs} \tag{32}$$

where $p_i^{rs} = c_i^{rs} + c_i^r + \bar{\pi}_i^r$, and where $\bar{\pi}_i^r$ is the mark up level in (22) wich reflects the same profitability (or rate of return criterion) in region r. By analogy with (21) we assume that delivery changes respond to E_i^s as follows:

$$\dot{x}_i^{rs} / x_i^{rs} = \alpha_i^{rs}\, \tilde{E}_i^{rs}$$

$$\tilde{E}_i^{rs} = \begin{cases} E_i^s \text{ if } \alpha_i^{rs} > 0 \\ 1 \text{ if } \alpha_i^{rs} \leq 0 \end{cases} \tag{33}$$

where $\alpha_i^{rs} = \alpha_i^r(\pi_i^{rs})$ is less than zero as $\pi_i^{rs} < 0$. In general we may think of α_i^r as a function with the following properties (compare Mansfield, 1961; Peterka, 1977):

$$\alpha_i^r(\pi_i^{rs}) > 0 \text{ as } \pi_i^{rs} \geq \bar{\pi}_i^r$$

$$\alpha_i^r(\bar{\pi}_i^r) > \alpha_i^r(\pi_i^{rs}) > 0 \quad \text{as} \quad 0 \leq \pi_i^{rs} < \bar{\pi}_i^r \tag{34}$$

$$\alpha_i^r(\pi_i^{rs}) < 0 \text{ as } \pi_i^{rs} < 0$$

Formulas (33) - (34) imply that x_i^{rs} is reduced when the link profit becomes negative, irrespective of which sign E_i^s currently assumes. With this formulation the competition between "identical" flows x_i^{ks} and x_i^{rs} will differ from that between x_i^{ks} and x_j^{rs}, $j \neq i$. In the latter the substitution is determined by α_i^{ks} and α_j^{rs} as well as by $A_i^s(p^s)$ and $A_j^s(p^s)$, where

$$A_i^s(p^s) = v_i^s(p^s)\, m^s/p_i^s = [\, 1 - \sum_j v_j^s(p^s)]\, m^s/p_i^s.$$

However, the substitution between x_i^{ks} and x_i^{rs} is not influenced by $v_i^s(p^s)$, since this factor is shared by the two competing flows. Their substitution is only affected by the difference between α_i^{rs} and α_i^{ks}. When $v_i^s(p^s)$ changes, the size of their joint market in region s will expand or contract.

4.2 Interaction Between Price and Quantity Adjustments

From (32) we can see that the development of p_i^s is affected by changes in both p_i^{rs} and x_i^{rs}. With two suppliers to region s, say 1 and 2, note that $\dot{p}_i^s < 0$ if $p_i^{1s} < p_i^{2s}$ and $\dot{x}_i^{1s} > \dot{x}_i^{2s}$. Suppose now that $\partial\alpha_i^r / \partial\pi_i^{rs} > 0$. Then the rate of growth will be faster for flows x_i^{rs} which have a cost advantage in the sense that $p_i^{rs} < p_i^s$. These are the flows with the highest link profits π_i^{rs}. The share $f_i^{rs} = x_i^{rs} / \sum_r x_i^{rs}$ will naturally grow for flows such that $\alpha_i^{rs} > 0$ if $\sum x_i^{rs}$ declines. The latter may occur if many suppliers have $\alpha_i^{ks} < 0$.

Consider now a situation in which $\dot{p}_j^s \geq 0$ for all $j \neq i$. Next, observe that

$$\dot{p}_i^s \sum_r x_i^{rs} = \sum_r \dot{p}_i^{rs} x_i^{rs} + \sum_r (p_i^{rs} - p_i^s)\, \dot{x}_i^{rs} \tag{35}$$

We know that flows which expand have $p_i^{rs} < p_i^s$. Hence, the last term on the right hand side is never positive. This implies that p_i^s is gradually reduced if $\sum \dot{p}_i^{rs} x_i^{rs} \leq 0$ and $\alpha_i^{rs} E_i^s > 0$ for at least one pair (r,s). When this applies we have $m^s \dot{f}_i^s \geq 0$, where $f_i^s = p_i^s \sum_r x_i^{rs} / m^s$. This means that $p_i^s \sum_r \dot{x}_i^{rs} > - \dot{p}_i^s \sum_r x_i^{rs}$. As long as this condition lasts there will be at least one flow for which $\dot{f}_i^{rs} > 0$ until product i is only delivered from region r, (i.e., $f_i^{rs} = 1$).

By analogy with (24), we can write $F_i^s = f_i^s / (1 - f_i^s)$ and $\dot{F}_i^s / F_i^s = \dot{f}_i^s / f_i^s (1 - f_i^s)$. Hence, we may argue in a similar way to the non-spatial case. Observe first that $\dot{f}_i^s m^s = \dot{p}_i^s \sum x_i^{rs} + p_i^s \sum \dot{x}_i^{rs}$. Introduce next a price development process of the same type as in (26), and approximate this process by

$$\dot{p}_i^s = k_i^s G_i^s (d_i^s) \dot{d}_i^s \tag{36}$$

where $d_i^s = \sum_r x_i^{rs}$. Then we may write (compare 28):

$$m^s \dot{f}_i^s = - k_i^s G_i^s (d_i^s) \dot{d}_i^s d_i^s + p_i^s \sum_r \alpha_i^{rs} x_i^{rs} \tilde{E}_i^{rs} \tag{37}$$

In this case the change in $G_i(d_i^s)$ will reflect the development of each cost term $c_i^{rs} = c_i^{rs}(x_i^{rs})$ and $c_i^r = c_i^r (\sum_s x_i^{rs})$. If the cost functions decline as the scale increases, $\dot{d}_i^s > 0$ will bring about a gradual fall in p_i^s. If we assume similar properties of G_i^s as G_i in (26), $G_i^s (d_i^s)$ will gradually approach zero if d_i^s does not become constant at an earlier stage. In either case the first term on the right hand side will approach zero. As the price level stagnates the second term will also be reduced because of the properties of E_i^s.

In summary, the previous characterization of the development of f_i^{rs} implies that the number of regions which supply identical products to each region s will normally decline. Each product group will therefore tend to contain only differentiated products. In this respect, the spatial model becomes very similar to the non-spatial one, and will likewise display similar properties. These properties include expansion and decline of various products along sigmoid paths. In the spatial case, abrupt discontinuities (catastrophes) will be more frequent due to the coexistence of basically identical flows heading for the same destination.

4.3 Spatial Product Cycles

One explanation for a gradual reduction of p_i^s over time is the emergence of a new supplier in the proximity of region s with lower delivery costs than the established suppliers. If the production technique has matured to such an extent that $c_i^r \approx c_i^k$ in many locations r and k, the establishment of production in a location k closer to the buyer (i.e. where $c_i^{ks} < c_i^{rs}$) will result in a lower value of p_i^{ks} than p_i^{rs} in the expression $p_i^s = \sum_r p_i^{rs} x_i^{rs} / \sum_r x_i^{rs}$. This type of spatial redistribution is rather typical as a market strategy for a mature product (see, e.g. Johansson and Karlsson, 1987; Norton and Rees, 1979).

The spatial model in section 4 has specific implications for the initiation of production in any region r. From (34) we know that deliveries will be directed to destinations where π_i^{rs} is higher than elsewhere. In general we may assume that $c_i^{rr} < c_i^{rs}$ for each $s \neq r$. If $p_i^r \approx p_i^s$, this means that the delivieries from producers in region r (i.e. x_i^r) will initially

expand more rapidly on the intraregional (r,r) than on any interregional link (r,s), extending from r. As production expands and production costs are reduced, the interregional deliveries may in a second phase increase very quickly.

As the relative factor costs, c_i^k, of other regions k begin to fall over time, the link-related profitability criterion (31) ensures that locations outside r may gradually become more favoured production centres; so much so that interregional deliveries then become more attractive. These spatial delivery patterns may be written down in the form of link-specific shares, viz:

$$h_i^{rs} = p_i^s x_i^{rs} / \sum_j p_j^s x_j^{rs} \tag{38}$$

The dynamic process of substitution over space may then be investigated by comparing, among others, the trajectories of h_i^r and h_i^{rs}. In order to describe the relation between substitution on the domestic and export markets we introduce the following variables

$$\begin{aligned}
f_i^r &= p_i^r \sum_s x_i^{sr} / m^r \\
h_i^{re} &= \sum_{s\neq r} p_i^s x_i^{sr} / \sum_{s\neq r} \sum_j p_j^s x_j^{rs} \\
p_I^r &= \sum_{j\neq i} \sum_s p_j^r x_j^{sr} / \sum_{j\neq i} \sum_s x_j^{sr} \\
p_I^{re} &= \sum_{j\neq i} \sum_{s\neq r} p_j^s x_j^{rs} / \sum_{j\neq i} \sum_{s\neq r} x_j^{rs}
\end{aligned} \tag{39}$$

Let $F^1 = f_i^r/(1 - f_i^r)$ and $F^2 = h_i^{re}/(1 - h_i^{re})$. For each of these F-variables we have tested a model with the following form:

$$\log F = a + bt + c(p_i/p_I) \tag{40}$$

where a, b, c are constants of regression, t denotes time, and (p_i/p_I) refers to (p_i^r/p_I^r) when $F = F^1$ and to (p_i^{re}/p_I^{re}) when $F = F^2$.

With regard to the domestic market, represented by F^1, preliminary investigations show that equations like (38) cover most of the dynamics in the market place (Batten and Johansson, 1985a; Johansson and Larsson, 1985; Larsson, 1985). These studies indicate that the development of relative prices (the term p_i/p_I) has a significant influence on the dynamics. At the same time, it is found that the pace of the substitution process frequently is quite steady. This corresponds to markets in which the relative price indicator p_i/p_I changes monotonically, as indicated in the formulation of (25) - (27). As a consequence, the substitution process can often be modelled as $\log F = a + bt$, where time acts as a driving force. Similar observations apply to the export market. In this case we illustrate our earlier suggestion that initially the intraregional deliveries (x_i^{rr}) dominate, while export

deliveries at a later phase may start to increase at a faster rate. Figures 3 and 4 depict typical results in the case of Sweden.

Observe that the logarithm of the ratio of market shares for each product group, namely log F, follows a straight line in both the domestic and export markets. Hence, link-specific product flows display the same type of development pattern as has been demonstrated in a non-spatial context (Fisher and Pry, 1971; Peterka, 1977; Batten and Johansson, 1985a). A noteworthy feature of the dynamics depicted in Figures 3 and 4 is that in the initial phase, $t < t_0$, $F^1(t) > F^2(t)$, which means expansion starts in the domestic market as suggested earlier. For the two examples t_0 is characterized by $f_i^r(t_0) \approx 0.4$.

At the same time we observe that $\dot{F}_2 > \dot{F}_1$, which should be a typical situation for all delivering regions such that the domestic market is significantly smaller than the nearby export markets. In the initial stages, low intraregional delivery costs favour local sales. However, the size of foreign markets and the strong impact of high delivery costs for declining products makes the substitution rate faster on the export links.

The above observations report on spatio-temporal patterns which may be referred to as spatial product cycles. As a complement to (38) and (39), it is also instructive to investigate directly the changing relationship between various interregional deliveries of the same product. We may define the following product-specific share

$$g_i^{rs} = p_i^s x_i^{rs} / \sum_s p_i^s x_i^{rs} \tag{41}$$

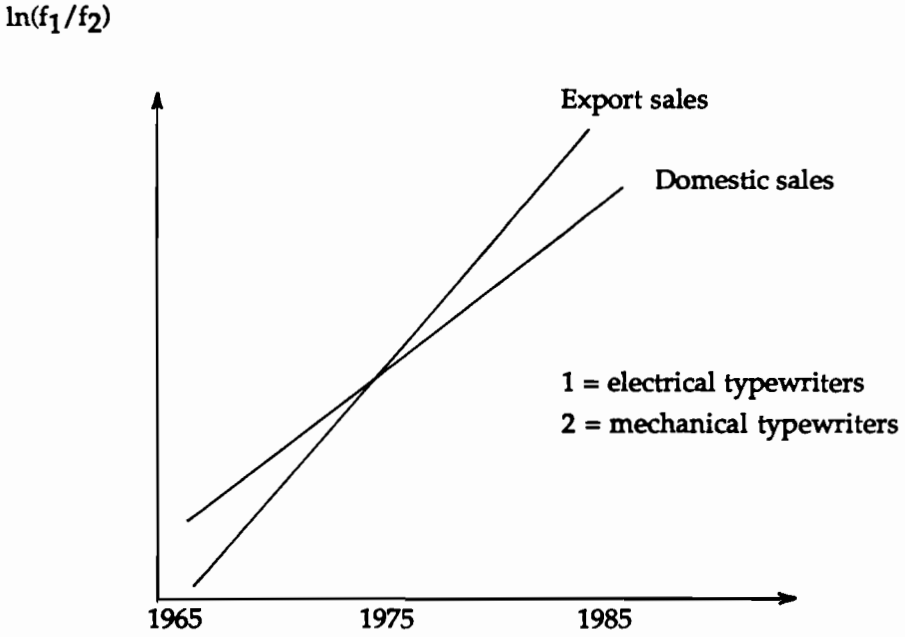


Figure 3 Substitution between electrical and mechanical typewriters

and then explore the temporal pattern of spatial substitution between g_i^{rk} and g_i^{rs}, constrained by $\sum_s g_i^{rs} = 1$ (i.e. $\sum x_i^{rs} = x_i^r$) and $\sum_s \dot{g}_i^{rs} = 0$ (i.e. $\sum \dot{x}_i^{rs} = \dot{x}^r$). An interesting example of this pattern is depicted in Figure 5, in so much as it displays the discontinuous changes which are predicted by our model whenever a subset of b^i-vectors in a product group are linearly dependent. The figure portrays the deliveries on export links from Sweden (in this case certain panel products). In this instance there is an abrupt fall in Sweden's delivery to Denmark of a kind we may call a "link catastrophe". Such a shift may (in the model) be caused by three interrelated forces which are described in Sections 3.3 and 3.4. One such force is competition from other differentiated products which generate a catastrophic drop in Denmark's budget share for the panel product in question. A related force is the competition from identical products supplied from other regions. A third cause for an abrupt change is the link profit condition, specified in (33) - (34).

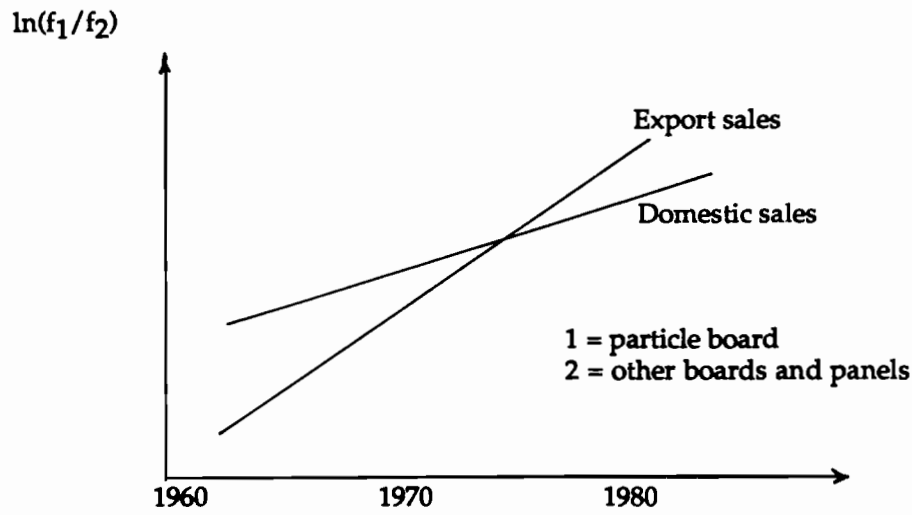


Figure 4 Market penetration of particle board

5. CONCLUDING REMARKS

The abovementioned substitution model for deliveries over space may be seen as a "traditional" extension of the general substitution model described earlier in this paper. It is traditional in the sense that space has been introduced by alluding directly to a set of market locations and to the transportation costs accruing from the resulting delivery problem. This view of the spatial economy is a standard one among regional economists.

There is another means of introducing space which would also be compatible with the general substitution model, but perhaps be more in keeping with its demand-based nature.

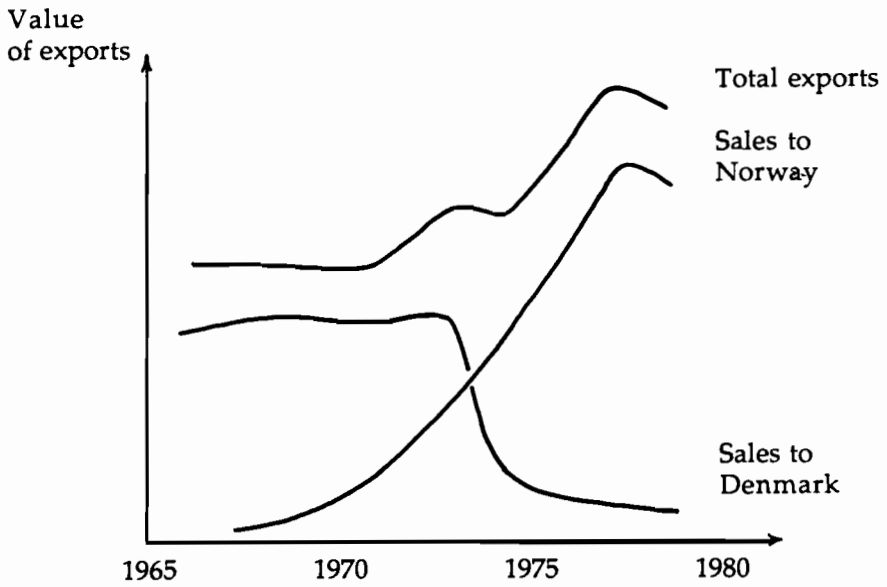


Figure 5 Swedish exports of plaster panels by region of destination

This would be to define the region of production as a basic characteristic k to be included as a component z_k, in each vector of characteristics z. The rationale for this lies with the empirical observation that the preference structure of certain customer groups is sometimes *biased* towards products manufactured locally, and often favouring supply from distinct delivery regions. Quite naturally, such preference structures vary for different types of customers and also may change over time. There are indications that, among some groups, the traditional preference for locally-made products - attributable at least in part to lower transportation costs - is gradually diminishing. Explanations may be traced to price competition and the proliferation of multi-national firms. In any event, it may be prudent and rewarding to include the region of origin in the vector of attributes, and to analyse the resulting trajectories. Such an approach falls within the analytical bounds of the general substitution model presented in Sections 2 and 3, and might therefore be explored as a spatial corollary to the non-spatial case. As a consequence, the excess demand expression $E_i^s = A_i^s(p^s) - \sum_r x_i^{rs}$ in (30) would change to

$$E_i^{rs} = A_i^{rs}(p^s) - x_i^{rs} \tag{42}$$

This version of spatial excess demand is a direct extension of the non-spatial formulation in (20). Adopting (42), all results in Section 3 would hold also in the inter-regional formulation. At the same time, this new model will be extremely sensitive to such changes in the preference structure which imply that the customers become indifferent with regard to the origin characteristic. Any change of this type over time would increase the probability of discontinuities and link catastrophes dramatically.

APPENDIX

A.1 Lemma 2

Consider a case with a single customer group, and omit the customer group index. From Lemma 1 we have that $F(p,m) = \{x \in K(p,m) : u(Bx) = u^{max}\}$ is upper semi- continous, given the assumptions about u and B in Lemma 1. These assumptions also imply that F is continuous when it is vector valued (i.e. not set valued). We shall show that F is vector valued and continuous when the column vectors, b^k, of B are linearly independent.

Let there be n goods, and let $z^k(p) = b^k m/p_k$. Moreover, let $z = z(p) = \sum v_k(p)\, z^k(p)$, where $\sum v_k(p) = 1$. Lemma 2 assumes that (i) u in (6) is strictly quasi-concave, continuous, increasing and differentiable, and (ii) B in (6) is an irreducible matrix and that the b^k-vectors are linearly independent. For $z = Bx$ and $z' = Bx'$, this implies

$$z \neq z' \iff x \neq x'$$

This means that $\phi(x) = u(Bx)$ will be a continuous, differentiable and strictly quasi-concave function. Hence, we can replace the optimization procedure in (8) by the (strictly quasi-concave) Lagrange function $L(p) = \phi(x) + \lambda(m - px)$, where $\lambda > 0$ and $px = \sum p_i x_i$. The properties of ϕ guarantee that the maximization of L yields a unique solution x for each p. By the implicit function theorem the solutions can be expressed as generated by a continuous function $x = x(p)$.

The above outline suffices to ascertain that $x = F(p,m)$ is continuous. Since $x_k = v_k(p)m/p_k$, the v_k-functions will also be continuous in p.

We shall clarify the above result further by considering a subset of products, 1, ..., r ($r \leq n$). The associated b^k-vectors are linearly dependent if and only if

$$z^1(p) = \sum_{k=2}^{r} \alpha_k z^k(p) \qquad (a.1)$$

where at least one scalar $\alpha_k \neq 0$. The properties of u and B imply that F is set valued only if one for some price p can find two constellations $v_1, ..., v_r$ and $v_1^*, ..., v_r^*$ such that

$$\sum_{i=1}^{r} v_i(p)\, z^k(p) = \sum_{i=1}^{r} v_i^*(p)\, z^k(p) \qquad (a.2)$$

as $v_i \geq 0$ and $(v_i + ... + v_n) = 1$. This means that F is set valued if, with an appropriate indexation of the products, we can find a p such that M and N in (a.3) satisfy M = N.

$$M = \sum_{k=1}^{r} v_k(p)\, z^k(p)$$

$$N = \sum_{k=2}^{r} v_k(p)\, z^k(p) + v_1(p) \sum_{k=2}^{r} \alpha_k\, z^k(p) \qquad (a.3)$$

where the α_k's apply to (a.1). From (a.1) follows that $M = N$ only if linear dependence applies. We may note that the v-coefficients associated with N can be expressed as $v_k^* = v_k + \alpha_k v_1$ for $k \neq 1$, and $v_1^* = 0$. Hence, with linear independence there is only one v-vector associated with each p.

A.2 Proposition 2 and Remark 2

The assumptions in Proposition 2 are the same as in Lemma 2. In addition we introduce the condition that there exists a subgroup of products $k \in I$ such that

$$b^1 = \hat{\alpha}_2\, b^2 + \ldots + \hat{\alpha}_r\, b^r\,, \quad r \leq n \qquad (a.4)$$

where all $b^k \geq 0$, $b^k \neq 0$, and all $\hat{\alpha}_k > 0$. By selecting a suitable set of prices we can see that (a.1) is just a reformulation of (a.4). The expression in (a.1) is transformed to (a.4) by choosing the α^k-coefficients in (a.1) such that

$$\alpha_k = \hat{\alpha}_k\, p_k/p_1 \qquad (a.5)$$

This follows from $z^k(p)/m = b^k/p_k$. We also observe that (a.1) - (a.4) concern one single customer group which we now select to be group g in Proposition 2. All this means that we can use the result in (a.3). Since b^1 contains all relevant attributes, we know that at the price p^* we can express M in (a.3) as $M(p^*) = b^1 m/p_1^* = z^1(p^*)$.Moreover, $v_1(p^*) = 1$, $v_k(p^*) = 0$ and $x_1(p^*) = m/p_1^*$. For N we get $v_1^* = 0$ and $v_k^*(p^*) = \alpha_k\, v_1(p^*) = \alpha_k$, where $\alpha_k = \hat{\alpha}_k\, p_k^*/p_1^*)$. By increasing p_1^* to $p_1 = p_1^* + \varepsilon$ we obtain (for $\varepsilon > 0$ and $p = (p_1\,, p_{(1)})$

$$\begin{aligned} &M(p_1^* + \varepsilon\,, p_{(1)}^*) < M(p^*) \\ &N(p_1^* + \varepsilon\,, p_{(1)}^*) = N(p^*) \end{aligned} \qquad (a.6)$$

Analogously, reducing p_1^* to $p_1^* - \varepsilon$ yields

$$M(p_1^* - \varepsilon\,, p_{(1)}^*) > M(p^*) = N(p_1^* - \varepsilon\,, p_{(1)}^*) \qquad (a.7)$$

Hence, (a.6) implies that $v_1(p_1^* + \varepsilon\,, p_{(1)}^*) = 0$, (a.7) implies $v_1(p_1^* - \varepsilon\,, p_{(1)}^*) = 1$, and $M(p^*) = N(p^*)$ implies that $v_1(p^*)$ is undecided. Noting that $E_i = mv_i(p)/p_i - x_i$, this completes the proof of Proposition 2.

The assumptions in Remark 2 mean that we cannot necessarily obtain the expression in (a.1). Instead we can obtain two groups $M_1 = M_2$

$$M_1 = \sum_{k=1}^{r} v_k(p^*)\, z^k(p^*)$$

$$M_2 = \sum_{k=1}^{h} \tilde{v}_k(p^*)\, z^k(p^*) \quad , \quad h < r \qquad \text{(a.8)}$$

where $\tilde{v}_1(p^*) > v_1(p^*) \geq 0$. This means that in M_2 at least one of the products 2, ..., r has left the convex combination. Therefore an increased budget share will go to product 1. In particular, this means that $v_1(p_1^* - \varepsilon\ ,\ p_{(1)}^*) - v_1(p_1^* + \varepsilon\ ,\ p_{(1)}^*)$ represents a discrete jump generated by an arbitrarily small $\varepsilon > 0$. This completes the proof.

REFERENCES

Berge, C.,1966, *Espace Topologiques: Fonctions Multivoques*, Dunod, Paris.

Batten, D.F.,1982, "On the Dynamics of Industrial Evolution", *Regional Science and Urban Economics*, 12:449-462.

Batten, D.F. and Johansson, B., 1985a, "Industrial Dynamics of the Building Sector: Product Cycles, Substitution and Trade Specialization", in F. Snickars, B. Johansson, and T.R. Lakshmanan, (eds.), *Economic Faces of the Building Sector*, Document D20:1985, Swedish Council for Building Research, Stockholm.

Batten, D.F. and B. Johansson, 1985b, "Price Adjustments and Multiregional Rigidities in the Analysis of World Trade", *Papers of the Regional Science Association*, 56:143-166.

Fisher, J.C. and R.F. Pry, 1971, "A Simple Substitution Model of Technological Change", *Technological Forecasting and Social Change*, 3:75-88.

Johansson, B., 1978, "Contributions to Sequential Analysis of Oligopolistic Competition, Memorandum 73, Department of Economics, University of Gothenburg.

Johansson, B. and J. Larsson, 1985, "Wood Product Industries and Building Materials", Research Report 1985:2, University of Karlstad (in Swedish).

Johansson, B. and C. Karlsson, 1987, "Processes of Industrial Change: Scale, Location and Type of Job", in M. Fischer and P. Nijkamp, (eds), *Regional Labour Market Analysis*, North-Holland, Amsterdam.

Lancaster, K., 1971, *Consumer Demand - A New Approach*, Columbia University Press, New York.

Lancaster, K., 1982, "Innovative Entry: Profit Hidden Beneath the Zero", *The Journal of Industrial Economics*, vol XXXI:41-56

Larsson, J., 1985, "Product Cycles, Substitution and Specialization", Research Report 1985:3, University of Karlstad (in Swedish).

Mansfield, E.,1961, "Technological Change and the Rate of Imitation", *Econometrica*, 29:741-766.

Norton, R. J. and Rees, 1979, "The Product Cycle and the Spatial Distribution of American Manufacturing", *Regional Studies*, 13:141-151.

Peterka, V., 1977, "Macrodynamics of Technological Change: Market Penetration by New Technologies", Research Report RR-77-22, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Advances in Spatial Theory and Dynamics
Å.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 4

Quality and Process Improvements in Dynamic Production Processes

J-P. Aubin

Åke Andersson has suggested that a distinction should be made between important components of technological change and R&D in the analysis of the dynamics of production processes. The first component, which we shall denote by u, embodies the strategy consisting of developing the quality of the products generated by a given technological process. The second component, denoted by v, describes a given technological process which has to be chosen from a class of available processes. An often observed evolution of producer's strategies consists of improving first the quality of the products for a given process and then, when this strategy is no longer viable, to look for a new technological process which allows the producers to satisfy the demand.

We propose in this paper to use viability theory for providing a mathematical metaphor for the idea that the evolution of technological process has more "inertia", or else, is "heavier", than the evolution of the quality of the products produced by a given technology. Viability theory deals with the study of dynamical systems - deterministic or not - whose trajectories must obey binding constraints called viability constraints. It does not involve optimization of intertemporal criteria, and thus, does not use optimal control techniques. This would assume the existence of one or several decision makers being able to control the evolution of the system, who agree over intertemporal criteria involving knowledge of the future (may it be stochastic) and who choose the optimal controls once and for all at the outset.

Instead, we are looking for controls which allow the trajectory of the system to remain viable. When choosing among such "viable" controls becomes a problem, we propose that the decision makers use *at each instant* mechanisms allowing them to minimize the *velocities of the controls*. Taking a decision such as this amounts to deciding in which directions and how fast the controls must change. The consequences of such decisions are discernible in the evolution of the trajectory.

We shall illustrate these points using a very simple metaphor of an entrepreneur producing commodity bundles x with quality characteristics described by a control vector u. For the sake of simplicity, we do not consider prices explicitly but treat them as one of the quality characteristics.

We assume that the aggregate demand for consuming such commodity bundles evolves according to a differential equation of the form

$$x'(t) = f(x(t),u(t)) \qquad (1)$$

which is controlled by quality characteristics u(t) of the commodity bundle x(t) at time t.

We assume that the production process is described by a set-valued map relating the given quality characteristics u and technological process v with the production set K(u,v) of commodity bundles x. In other words, the production process is represented by viability constraints of the form

$$x(t) \in K(u(t),v(t)). \tag{2}$$

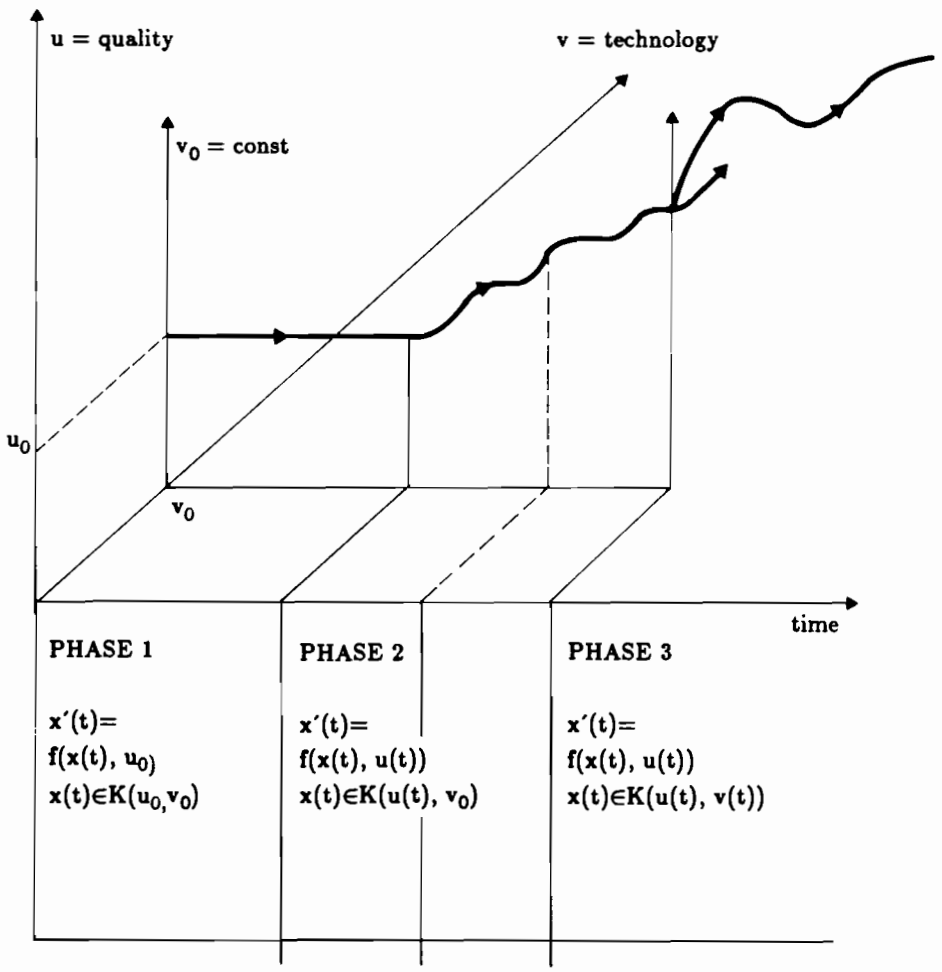


Figure 1 Three phases of the evolution

If we start at the initial instant t = 0 with a viable initial state

$$x_0 \in K(u_0,v_0) \tag{3}$$

the first problem is to obtain necessary and sufficient conditions linking the dynamical system (1) and the production constraints (2) which jointly specify viable trajectories of (1), (2) for every viable initial condition (3). Then we shall describe an evolution of such a system where we hold constant for as long as we can, first the technological process, then the quality characteristics; whereas the production evolves at will.

For instance, starting at time t=0 with u_0, v_0 and $x_0 \in K(u_0, v_0)$, we let the state evolve according to the differential equation $x'(t) = f(x(t), u_0)$ as long as $x(t)$ remains in $K(u_0, v_0)$ - phase 1. When this is no longer possible, the technological process remains constant, but the quality characteristics $u(t)$ start to evolve, and the evolution of the production is governed by the controlled system $x'(t) = f(x(t),u(t)$ (as long as $x(t)$ remains in $K(u(t),v_0)$ for all t (phase 2). When this is no longer possible, the producer must change his technological process and we enter phase 3. If at a later time the technological process can be kept constant, the system returns to phase 2 again from which it can evolve either to phase 3 or to phase 1, from which further phase changes of a similar kind are possible.

We can represent this type of evolution in a diagram (see Figure 1) where the evolution can be divided into three phases, each of which bears some relationship to the lifecycles of a product of the production process as discussed widely in the economics literature.

To achieve our stated goals, we need to introduce an adequate concept of the "derivative" of the set-valued map K describing the production process. This is not such a difficult task, since it suffices to return to Fermat's (1637) definition of a tangent to a function: If f is a differentiable function from R to R, the graph of the linear map $u \to f'(x)u$ is the tangent space to the graph of f at the point (x,f(x)) as shown in Figure 2.

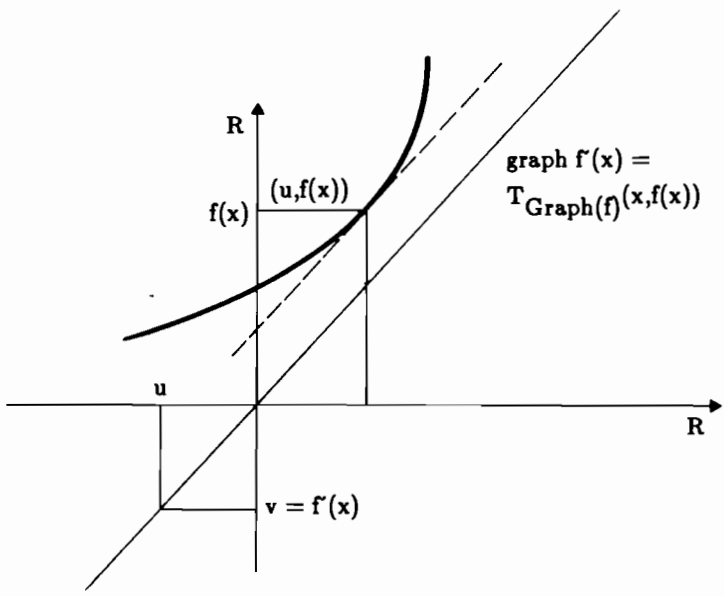


Figure 2 Graph of f at the point (x,f(x))

We adopt the same idea for a set-valued map F from a vector space X to a vector space Y. We take a point (x,f(x)) as demonstrated in Figure 2.

$$\text{Graph F: } = \{(x,y) \in X \times Y \mid y \in F(x)\}.$$

We cannot define the concept of "tangent space" to the graph of F at (x,y), which is an arbitrary set, without making very restrictive assumptions. But by relaxing our demand, and being content with "tangent cones" instead of tangent spaces, we can always define the concept of "contingent cone to the graph of F at (x,y)"

$$T_{\text{Graph }(F)}(x,y)$$

which is always a *closed cone* (instead of a vector space). Now, we can regard this cone as the graph of a new set-valued map (depending upon F, x and y) DF(x,y) from X to Y, defined by

$$v \in DF(x,y)(u) \iff (u,v) \in T_{\text{Graph}(F)}(x,y)$$

which may be called the "contingent derivative" (see Figure 3).

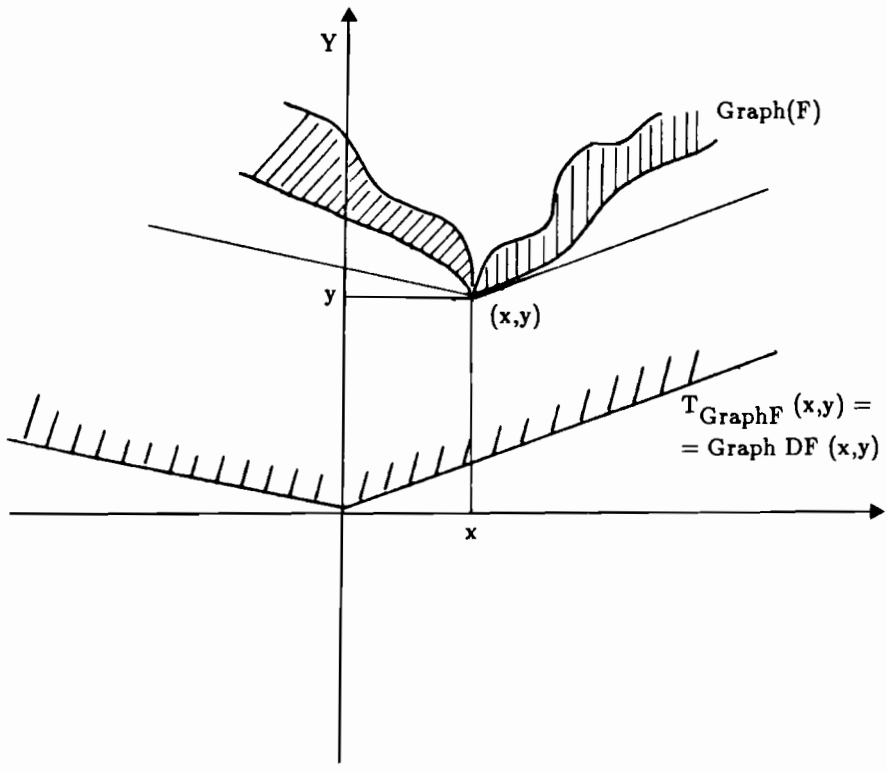


Figure 3 The "contingent derivative"

Hence the "contingent derivative" DF(x,y) of a set-valued map F at a point (x,y) of its graph is also a set-valued map F from X to Y, which has a closed graph (a weak extension of the concept of continuity) and which is positively homogeneous

$$\forall \lambda > 0,\ DF(x,y)(\lambda u) = \lambda DF(x,y)(u)$$

(a weak extension of the concept of linearity).

Needless to say, if F is a usual differentiable (single-valued) map, then the contingent derivative coincides with the usual Frechet derivative

$$DF(x,F(x))(u) = F'(x)u.$$

Many properties of derivatives of differentiable maps can be carried over to contingent derivatives of set-valued maps (for an account of non-smooth analysis, see chapter 7 of Aubin and Ekeland 1984).

Now consider the following "chain rule" which plays a crucial role in our study. If x(t) and y(t) are differentiable functions such that

$$y(t) \in F(x(t)) \text{ for all } t \geq 0.$$

then

$$y'(t) \in DF(x(t),y(t))(x'(t)) \text{ for all } t \geq 0.$$

We also see at once that the derivative of the inverse F^{-1} of F, defined by

$$x \in F^{-1}(y) \Leftrightarrow y \in F(x)$$

is the inverse of the derivative

$$D(F^{-1})(y,x) = DF(x,y)^{-1}.$$

This is a very useful formula, even when F is single-valued.

Then, by "formally" differentiating the viability conditions (2), and taking into account the differential equation (1), we obtain the formula

$$f(x(t),u(t)) = x'(t) \in DK(u(t),v(t),x(t))(u'(t),v'(t)).$$

We can invert this and derive the velocities of the evolution of viable quality characteristics and technological processes by way of the formula

$$(u'(t),v'(t)) \in D(K^{-1})(x(t),u(t),v(t))(f(x(t),u(t)). \tag{4}$$

In other words, we can derive from the original problem (1),(2) another "differential inclusion" governing the evolution of the controls u(t) and v(t) regulating viable trajectories.

We can also clarify the technical assumptions of a theorem stating in essence that the necessary and sufficient condition for a solution to (1), (2), (3), (4) to exist is that there exists some bound M such that B_M - denoting the ball of radius M - satisfies the following

$$\begin{aligned} &\forall (u,v,x) \in \text{Graph } K, \\ &f(x,u) \in DK(u,v,x)(B_M,B_M). \end{aligned} \tag{5}$$

This is the relation linking demand (whose evolution is described by the differential equation (1)) and production (described by the viability constraints (2)). (See Aubin and Cellina 1984 for precise statements in this respect).

Now assume that conditions (5) are satisfied. We know that we can solve the evolution of the commodity bundles, their quality characteristics and the technological process. Differential inclusion (4) shows that, in general, the producer has to choose at each instant velocities u'(t) and v'(t) with which he will change either the quality characteristics, or the technological process, or both.

We can naturally propose a variety of mechanisms allowing the producer to make such choices. Here we shall describe the evolution proposed by Å. Andersson.

For this purpose, assume that we are at time t and that x(t), u(t) and v(t) are known. We also know f(x(t),u(t)) and, in theory, the set

$$R(t) = D(K^{-1})(x(t),u(t),v(t))(f(x(t),u(t)).$$

We now have to choose the pair (u'(t),v'(t)) in the set R(t).

Case 1

If by any chance, the pair (0,0) belongs to R(t), this means that the zero velocities of the two controls are viable, so that the system can keep the controls u(t),v(t) constant at time t. In other words, for all times $s \geq t$, the system may keep the controls u(t),v(t) constant as long as the state x(s), which continues to evolve according to the differential equation

$$x'(s) = f(x(s),u(t))$$

remains in the "*viability niche*"

$$x(s) \in C(u(t),v(t))$$

where

$$C(u,v) := \{x \in K(u,v) \mid f(x,u) \in DK(x,u,v)(0,0)\}.$$

Thus, "heavy behaviour" of a producer implies that he will make this choice: keep constant both the quality characteristics and the process as long as the system remains viable.

Case 2

Assume instead that we are in a state when $v'(t) = 0$ and $u'(t) \neq 0$. Then the producer is permitted to keep his technological process constant, but is forced to change the quality characteristics of the commodities. To take into account that the evolution of the processes is "heavier" than the evolution of quality characteristics, we shall take $v'(t) = 0$ and keep the process constant as long as the system remains viable.

Case 3

Assume now that we are in a state when $u'(t) = 0$ and $v'(t) \neq 0$. The producer has the possibility of keeping the quality characteristics of the commodity constant and, for that

purpose, to change the process. However, if he obeys an evolution where the evolution of processes is heavier, he will not implement this choice: he will change both the process and the characteristics, taking the velocity v'(t) of the smallest norm and then, choosing u'(t) such that (u'(t),v'(t)) belongs to K(t).

General Case

In all cases, the producer will choose the pair of velocities (u',v') in R(t) by

a) first taking $\bar{v}'(t)$ such that

$$\|\bar{v}'(t)\| = \min \{ \|v'\| \mid \exists\, u' \text{ with } (u',v') \in R(t)\}$$

b) second, taking $\bar{u}'(t)$ such that

$$\|\bar{u}'(t)\| = \min \{ \|u'\| \mid (u',\bar{v}'(t)) \in R(t)\}.$$

This choice mechanism is quite easy to represent in a diagram (Figure 4). We represent in R^2 the points of coordinates
$(\|u'\|, \|v'\|)$ and by S(t): = = $\{(\|u'\|, \|v'\|) \mid (u',v') \in R(t)\}$.

We introduce the following sets

* V(x,u,v) is the set of velocities $\bar{v}'$ of minimal norm among velocities v' such that

$$(u',v') \in D(K^{-1})(x,u,v)(f(x,u)) \text{ for some } u'$$

* U(x,u,v,v') is the set of velocities $\bar{u}'$ of minimal norm among velocities u' such that

$$(u',v') \in D(K^{-1})(x,u,v)(f(x,u)).$$

Then the system of differential inclusions governing the evolution of viable solutions of (1), (2) when the evolution of technological process is heavier than the evolution of quality characteristics is

$$\begin{aligned} &\text{i)} \quad x'(t) = f(x(t),u(t)) \\ &\text{ii)} \quad v'(t) \in V(x(t),u(t),v(t)) \\ &\text{iii)} \quad u'(t) \in U(x(t),u(t),v(t),v'(t)). \end{aligned} \tag{6}$$

Such differential inclusions can be solved under "reasonable" technical assumptions using the methodology of Aubin and Frankowska (1985).

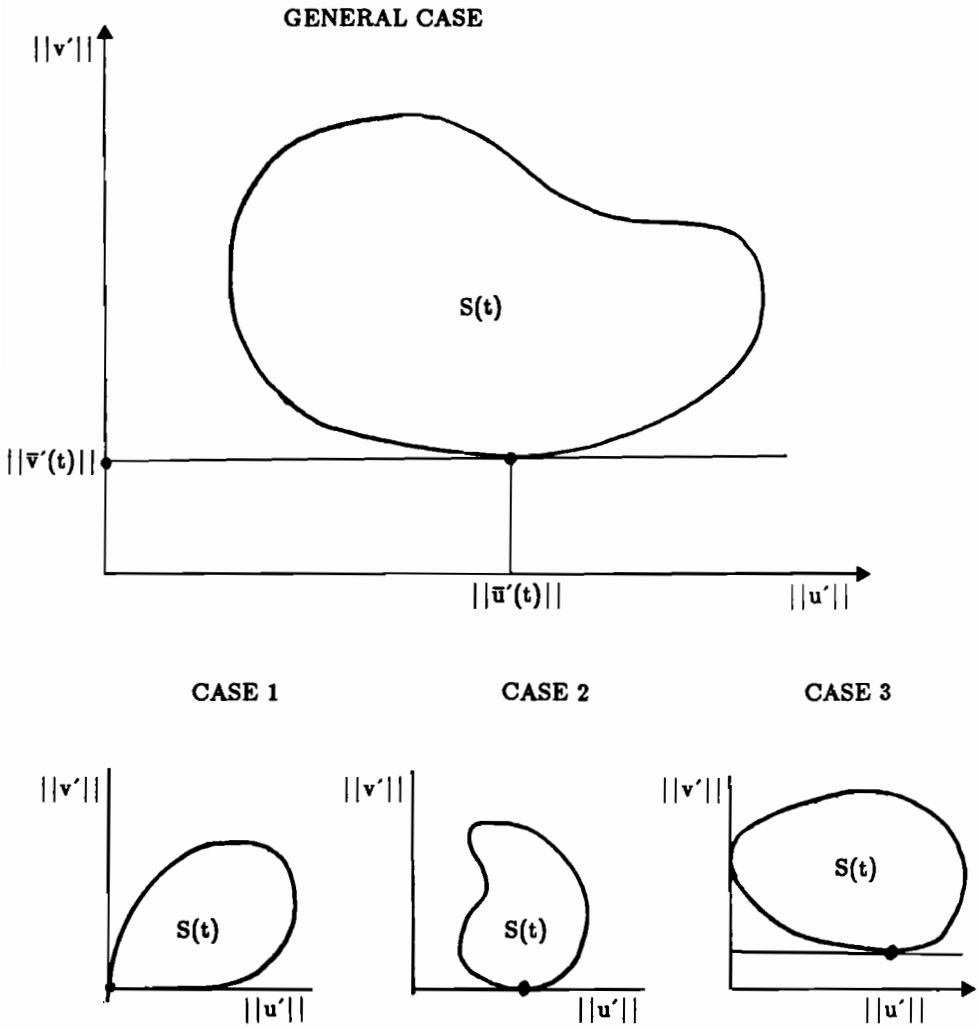


Figure 4 The choice mechanism

REFERENCES

Aubin, J-P. and A. Cellina, 1984, *Differential Inclusions*, Springer Verlag, Berlin.

Aubin, J-P. and I. Ekeland, 1984, *Applied Nonlinear Analysis*, Wiley Interscience, New York.

Aubin, J-P. and H.Frankowska, 1985, "Heavy viable trajectories", *Analyse Nonlinéaire*, Annales de L'Institut Henri Poincaré, Paris.

Advances in Spatial Theory and Dynamics
A.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 5

Women and Technological Development

L. Chatterjee

1. INTRODUCTION

"The work of a woman is never complete" is a common folk saying in various cultures. While most of the early research on women's work was related to household sustenance and reproduction, interest in the income generation activities of women is now increasing. During the last decade an extensive volume of literature has been produced on the labor of women - for the household and for the market. This paper is in this genre and focuses on the association between technological developments in production processes and the paid and unpaid work of women. It argues that women have facilitated innovations in technology and that this aspect has been largely ignored in the literature. Most attention has been paid to the impacts of technology on women rather than the way in which women have fostered technological developments in production and consumption processes.

The relationship between technological change - as represented by newer or improved products, processes, production structures and organizations of production and consumption - and gender remains complicated. There is a contradictory relationship between gender and technology in that technological developments have had both negative and positive impacts on the laboring characteristics of women. The history of women and technology reflects a series of occupational readjustments which have helped women to become emancipated from unpaid production and dependent relations at home and allowed them to obtain paid employment in the market. Yet this transformation has been accompanied by great hardships and onerous costs that have involved exploitative wages, tedious repetitive work on machines designed for the greatest productive efficiency, occupational health hazards and constant ideological struggles to justify the right for women to work in technologically improved work environments.

A major strand of research about the interface of women and technology documents the negative impact of technology on the employment of women, particularly the displacements caused by the introduction of new technology (Dauber and Cain, 1981; Knight and Wilmott, 1986; Maillier and Rosser, 1987). Technology, particularly that related to the operation of machines, commonly decreases the labor component in any production process and this is followed by a decline in employment opportunities in the technologically obsolete sectors. Labor historians and labor market analysts have shown that women are often concentrated in sectors that are technologically weak (Berg et al., 1983; Hanawalt , 1986). Consequently, the introduction of technology in these sectors has had a disproportionate impact on women - leading to the lowering of their wages through labor market imbalances. An oversupply of workers relative to shrinking demand has led to the impoverization of women workers and their households. In such cases technological developments have marginalised women.

The introduction of technology, however, has also opened up avenues for the employment of women. On the aggregate, technology increases the demand for labor since new products are developed and new needs are created. In the past, new products have often freed women from labor intensive, unremunerated labor in the home and increased the range of possible employment opportunities in the market. Also, working conditions and wages have improved with the increase in the range of work opportunities, at least relative to those existing for women in earlier historical periods. However, men have improved their working conditions and increased their wages more than women through these technologically induced advancements in the manufacturing and service sectors. There has been an unequal gender appropriation of the benefits of technological advancement, even as women have fought for and achieved greater emancipation through technological developments.

There is considerable empirical evidence from developing and advanced economies about these dual impacts of technology, though the primary research emphasis has been on documenting the negative impacts of technology on women's labor. Relatively little attention has been paid to the impacts that women have had on the development of technology. This paper briefly reviews the literature concerning impacts of technology on women and argues that the empirically observable impacts can be more usefully interpreted in the context of their functional relationship with the promotion of technological developments in society. This paper focuses on how women, through their adjustments to positive and negative technological impacts, have furthered the development and incorporation of new technologies in the production process. That is, women have not been the passive recipients of negative impacts of new developments of technology; rather, through their active adjustments to technologically induced displacements in their activities, they have aided the progression of technological developments in society. In this facilitative role they have borne more than their "fair share" of the costs of wastage and breakage associated with technological progress.

This paper adopts the increasingly dominant feminist perspective that there is a close interaction between the two domains of women's work - their unpaid work for the sustenance and reproduction of the household and their income earning activities. Both these domains are linked through the time budget. An analytical separation of women's work for the market and women's work for the household is theoretically unwarranted, particularly in the analysis of women's role in facilitating technical change. Many technological advancements have been stimulated in order to increase the allocation of women's work time in favor of work for the market. Household labor saving products have stimulated inventions and technological developments in infrastructure and appliance industries. Since women have had to combine the time demands of two work domains - income earning work and household maintenance work - this has important ramifications for the role of women in facilitating technical change. In the contemporary period new adjustments in women's production of services can be explained through the concept of coproduction. Coproduction and its relation to service production in the two work domains will be explored later in the paper. However, women have always had to divide their labor time between the two productive spheres. This division is intricately related to technological developments in production and consumption in society.

The paper is divided into four sections. There is a conceptual overview that discusses and very briefly summarises the findings of the rich literature about women and their laboring characteristics. The predominant attributes of female employment are linked to the progression of technology in various sectors of the economy. In the two following sections, the paper discusses the role of women in promoting technical change in the manufacturing and the service sectors. In the concluding section, a framework for analysing the role of women in technical change is provided.

2. CONCEPTUAL OVERVIEW

In this section a general statement is made about the relation between technology and women's employment. The relationship between technological change and employment has received and continues to receive considerable analytical attention. In this literature, however, the issue of gender has received less attention. Yet, technological developments have increased and continue to increase women's employment opportunities; in most countries of the world the percentage of employed women has increased during the post war years (ILO, 1981; OECD, 1984). For example, in the U.S. women accounted for 16% of the labor force in 1870; by 1986 their share had increased to 44% (Bergman, 1986). Urquhart (1984) has shown that the primary source of new employment in most advanced economies has been the employment of women who did not previously hold jobs. Indeed, female employment has been increasing at a faster rate than that of males in most high income countries (Chatterjee and Lakshmanan, 1987; Maillier and Rosser, 1987) and this also holds for less developed countries. In the majority of third world countries, the number of women in formal wage employment has increased more rapidly than that of men in the last two decades (UNIDO, 1980). The industrial sector has absorbed a major segment of this formal labor force, particularly in the manufacture of textiles, clothing and electronics. Nevertheless, the role that the market participation of women plays in technological change remains inadequately explored in the vast contemporary literature on technical change.

It must be evident, even to the casual observer, that women at all stages of the life cycle are drawn into the labor market in great numbers when major technical changes occur in the market place. The expansion of women's employment during periods of technological change is not only a contemporary phenomenon. There is considerable evidence that, even in the proto-industrial phase, emerging types of manufacturing production made special use of a female labor force, adapting types of technical change specially suited to the work patterns of a gender specific labor force. For example, economic historians have drawn attention to the fact that technical change in the textile industries in the years between pre-industrial and industrial 18th Century Britain required and adapted to the large scale use of female labor (Berg et al., 1983; Berg, 1985).

Technical change not only meant mechanization, but also the use of intermediate techniques, changes in work organization, wider use of inexpensive labor, an increasing division of labor, and variable industrial structures. What are the characteristics of a gender bias in technological development? What role did the household production of women play in determining the structure of work organization and the response of women to the new technologies? An understanding of this complex women-market-household production nexus is crucial in unravelling the role women play in promoting technological change.

Three existing strands of research are relevant to an understanding of this nexus. These three research themes are:

- Sex segregation in the workplace
- Wage inequalities between male and female workers
- Ideology of the supplementality/marginality of women's work and its relation to patriarchial elements of society.

The division of labor based on gender has been and is still common in both advanced and developing economies. Reskin (1984, p.2) notes there are three expressions of this sex segregation. First, women are relegated to different spheres of activity - women to the home activities and men to market activities. Second, paid employees work in predominantly single-sex settings. In the U.S. in 1980, over 32 million workers were employed in industries where the workforce was at least 80% male or female. The metal, construction and railroad industries are more than 90% male. Women constitute more than 75% of the workers in direct sales, apparel and fabricated textile manufacturing. Third, there is a functional segregation even in those industries which are integrated - that is men

and women do different jobs within the same industry. In hospitals, for example, nurses and clerical workers are predominantly female and doctors and administrators are predominantly male. In spite of remarkable changes in the composition of the workforce and occupational structure, sex segregation levels have remained stable throughout the nineteenth and twentieth centuries. While some gains have been made in the sexual integration of jobs, most men and women work in sex typical jobs (Reskin and Hartmann, 1985). Women are concentrated in a small range of feminized jobs, most often in less skilled, lower paying jobs relative to their male counterparts with equivalent education and age.

In developing countries the manufacturing sector is also divided between feminized industries and those where women are not employed in significant numbers. Aggregate statistics on occupational sex segregation have been documented by the ILO (1981). In addition, several case studies have traced this for specific industries in a variety of countries in Asia, Africa and Latin America. For example, in Morocco in the three main industries which account for 75% of the manufactured exports - carpets, clothing and food processing - women account for 70-100% of the workforce (Joekes, 1985). That women have been incorporated into the industrial workforce on the basis of sex segregation has been documented for the offshore multinational electronics factories in Malaysia and Singapore, for textiles and assembly plants in Mexico, manual workers in Brazil (Humphrey, 1985), and urban manual and clerical workers in India (Joshi and Joshi, 1976). The adoption of technological inventions in the production process has been made possible due to sex segregation. The association between sex segregation and technological developments will be illustrated for the manufacturing and service sectors in a later section of this paper.

Two of the most consistent descriptors of female employment are the segregation of the sexes in the workplace and the gap between the earnings of men and women. These two variables are correlated. Sex segregation has aided the practice of paying low wages to women workers at various stages of the production of final goods and services. A wage gap has persisted since the early industrial revolution in the U.K., when women in increasing numbers began to work for wages outside their household. The wage gap has remained despite the extensive labor force participation of women and efforts of the women's movement. The 'average' pay of females, whether involved in manual or non-manual work, is always lower than that of males at each age level (Maillier and Rosser, 1987). For example, in 1987 in the U.S., more than 25 years after the women's movement made the wage gap an issue, women were earning 70 cents to the dollar, compared to what men earned (Mann and Hellwig, 1988). This can be compared to 60 cents to the dollar in the early sixties and 63 cents to the dollar in the late seventies. While the gap is being reduced at a slow rate, it is not primarily due to comparable wages for comparable jobs. Rather, this minor reduction reflects the loss of higher wage blue collar jobs by men and the selective migration of women into male preserves such as law, medicine, business and engineering. Treiman and Hartmann (1981) provide a comprehensive analysis of the characteristics of unequal wages. Roos and Reskin (1984) have focused on the separate job ladders and exclusionary practices that restrict mobility between male and female positions.

The wage gap exists in developing countries as well (Dauber and Cain, 1981). This is true not only in the low technology traditional sectors. For example, Joekes (1985, p.183) shows that in Moroccan clothing factories which use modern technology to produce for the European clothing market, women machinists earned about 70% of the pay of male machinists performing identical tasks on the assembly line with similar age, experience and hours of work. This wage gap facilitates the incorporation of technology into production processes since women provide a low wage reserve labor force that is important in periods of technological transformation. There are structural parallels between the increasing part time work of women in the areas of the service sector which are characterized by low wages, and the low wage incorporation of women in the

manufacturing sector. The connections between low wages, sex segregation and technology development shall be illustrated later.

Sex segregation and low wages for women, both of which are important for technological development, have been justified on ideological grounds. Matthaei (1982) discusses the undervaluation of women's work, particularly the concept of the gender hierarchy of value that has been used to justify lower wages for women. According to patriarchal ideology, women were expected to be temporary workers, gainfully employed in predetermined periods of their life cycle - between maturity and matrimony. Though the reality of women's participation in the labor force was quite different, this argument justified the lack of investment in and commitment for skill development amongst women. Women learnt their skills in an informal manner through practical experience and a trial and error method. Their lack of skills was used to justify lower wages, through a circular reasoning. In a sex segregated labor situation, jobs with lower skill, productivity and technology were available to women and the adaptability of women provided a major source of energy for the recomposition of economic activities during periods of technical change.

The ideology of supplementary wages still exists in the third world where industrial wages can be extremely low and insufficient for the subsistence needs of women. For example, Mather (1985, p. 158) discusses the low wages received by women in West Java. Even though female wages provided a major source of money for food, clothing, and shelter in working class families, there was a tacit agreement between parents and factory managers about the supplementarity of female wages. This permitted employers to pay wages which were below subsistence level and kept the women dependent.

During periods of technical change it is necessary to have a pliant and efficient labor force that is willing to accept new work situations. In the next two sections documentary evidence will be provided to demonstrate the adaptability and efficiency of women, both at the experimental new technology end and the residual, obsolescent end of the technology spectrum in the manufacturing and service industries. There is a need for low wage labor since in the former there is the risk minimization logic of the entrepreneur and in the latter there are in reality lower productivities.

The comparative level of women's wages has been an issue for feminists for more than a hundred years. Yet there has also been resistance from union organisers, as demonstrated by Feldberg (1983), who discussed the influential role of union organisers in fighting for higher wages for men relative to women. The ideology of skill and family wages was used in this fight. The concept of the family wage confirmed the perspective of women as supplementary wage earners and extended the sexual division of labor in the household to the market. Ideological arguments were also used to justify the part time status of women workers.

These three variables, which reflect the conjunction of ideology and practice with respect to women in a patriarchial society, can be directly linked to the role women have played in the incorporation of technology as innovations in the production process and their access to technological improvements. The differential pace of technological development and its adoption in production processes has been made possible through gender biases in the development and differential allocation of technological improvements. Technological advances are usually made earlier in sectors of production dominated by male workers, since their wages are higher and there is a greater incentive for capital labor substitutions in those sectors. When technological advances are made in the female dominated sectors this is commonly followed by displacement of female workers in favor of male workers. Higher wages are possible since productivity increases due to improvements in technology and male workers appropriate the higher wage jobs. Such jobs are reclassified as 'skilled' through this process of technological improvements. However, displacement of women workers does not eliminate the need for women's employment. They shift to other low technology, low productivity sectors which then become sex segregated and have low wages attached to them. In fact, it can be

argued that low wages for women and their consistent placement in low productive sectors have permitted the development of different production structures that maximise profits and returns to capital.

This pattern of displacement and emplacement of female workers and their unequal access to technological improvements can be found in all periods of history and in a wide spectrum of developed and developing countries. What explains this consistency and why have men and women accepted this unjust practice? The ideology of women's primary work as pertaining to household maintenance and reproduction and their work for the market as being supplementary has provided the rationale for this.

We can see this ideological debate with respect to women's employment in each cycle of technological development whether it is in agriculture, manufacturing or services. It is ironic that the clerical sector is dominated by women, yet three arguments were originally used against the entry of women into the clerical professions. It was argued that clerical work was not suited to women's nature, they lacked the capacity for physical endurance and that their entry would displace men. All three arguments were derived from patriarchial values that define femininity in terms of women's domestic role and the supplementarity of their incomes.

In the next two sections these concepts are explored with respect to the manufacturing and service sectors.

3. THE ROLE OF WOMEN IN THE DEVELOPMENT OF INDUSTRIAL TECHNOLOGIES

The role of women in the development of industrial technology results from the conjunction of the three factors mentioned earlier and the profit making needs of entrepreneurs. Historical evidence will be used in this section to demonstrate that functional segregation of the sexes, and the low wages of women, have been used for the advancement of technology in society. Functional segregation, as noted earlier, results from the classification of specific jobs as male and female within integrated industries. While it is possible to document the existence of functional sex segregation in job tasks at any point in time, there is continuous reallocation of tasks between men and women over time. Technological change in the industrial sector has been associated with the twin processes of dynamic incorporation of women into production processes that were male dominated and the incorporation of males into tasks that were female dominated. For example, in the U.S. women were actually the first industrial workers (Glazer, 1980, p. 264). They were displaced by males, primarily immigrants. Since union organizers viewed women workers as threats, the ideology of female attributes and of supplementary wages were used to deny women access to industrial work after their displacement. Until the First World War blue collar work was considered to be a male activity except in a few feminized industries. The enhanced demand for workers and the restricted labor supply created by war time needs allowed women to demonstrate that they were competent to work in engineering and other "male" jobs. The continuous reallocation of tasks between men and women implies that no task is intrinsically male or female. It is societal needs that determine what at any particular phase will be considered as specifically "women's tasks" and on occasion through redefinition, as predominantly "male tasks". If this observation is accurate then we need to explore the association between the varying expressions of functional segregation and technological change.

While the actual tasks performed by women and the machines they operated varied in time, their role in the development of technology in the production process has remained constant. Functional sex segregation permitted them to be employed in tasks embedded with lower levels of technology and low wages allowed profit extraction from their labors even though overall productivity could be low in activities using obsolescent technologies. During periods of technological change women were involved in both the

high risk experimental phase of new technology and in the high risk, unstable, obsolescent phase of old technologies. During the stable phase of an industry women worked in subsidiary and auxiliary jobs. They aided male operators of machines and made their use of new technological advancements more efficient - hence more productive. That is, through their performance of auxiliary tasks they not only facilitated the control of more productive technology by male workers, they allowed capital formation to proceed more efficiently - capital that could be reinvested as new technology. However, when new technology was introduced into female dominated sectors and productivity increased, the benefits of higher productivity were appropriated by male workers. These tasks became reclassified as male preserves.

However, two aspects remained constant. First, women were paid lower wages than men, irrespective of whether they were in the experimental phases of new technology, the stable phase of mature technology or the declining phase of obsolescent technology. Women's wages were not commensurate with those of men in the different phases of technological development in a particular production process. Second, sex segregation remained as a characteristic of the male-female division of work. Through continuously readjusting in their work spheres, women facilitated the process of technological advancement. What guided this reallocation of tasks? Why was this reallocation necessary? How was this process of a changing sexual division of labor related to the development of technology? In summary, maximum returns to capital from the existing technologies and investments of capital for new technology were facilitated through these apparently changing manifestations of functional sex segregation.

One consistent thread that ties together these shifting task responsibilities is the definition of what constitutes "skill". Skill, in the labor process, is not defined in terms of human attributes. The definition of skilled and unskilled has been based on productivity rather than on the true skills required for production. Since women have been primarily incorporated into sectors and tasks that had lower levels of technology embedded in them, or were technologically obsolete, they have had lower productivities. The use of lower technologies that caused 'lower productivities' permitted women to be classified as unskilled and to receive lower wages. In many instances lower technologies required higher levels of skills. On the other hand, when male workers were reallocated to female tasks with new production processes and higher levels of technology, these tasks were reclassified as 'skilled' since productivity was increased through the incorporation of new technology. Higher productivity permitted higher wages of which the men were beneficiaries.

It is short sighted to concentrate on one phase of the cycle - the displacement of women by technology and/or male workers - without paying attention to the whole cycle and the role of technology in this. While such practices are clear examples of sexual discrimination and exploitation that can be traced to patriarchial relations in society, there has been a functional reason for this that is related to the way in which technology is developed and incorporated into production processes. Through this redefinition of tasks as male and female, society permitted the gradual development and incorporation of technology, which is necessary for two purposes. First, technological advancements are made in a piecemeal way due to the idiosyncratic nature of the development of inventions. Second, there are economic risks attached to the transition from inventions to innovations. Industrialists seek to minimise their risks by accepting innovations that have been tried and proven successful by a few entrepreneurs. The piecemeal introduction of technology into production processes has possibly been due to the adaptive role that women have played in accepting jobs and the lower wages provided by them in a society dominated by patriarchial values. Since the process of the invention of technology and the adoption of inventions into the production process as innovations is uneven, women - through their acceptance of lower wages, residual employment and absorption of the negative impacts of technological development - have permitted entrepreneurs to selectively adopt technological innovations into their production processes. Their lower wages have

allowed entrepreneurs to experiment with technological developments and to continue with the replacement of obsolete technology at a pace suited to their profit making needs. The ideology of supplemental earnings, weaker physical stamina and like values that were used to justify the lower wages and sex segregations were external manifestations of deeper structural forces that relate to the relation between labor and capital as defined in a patriarchial context.

These patriarchial ideologies facilitated the adoption of technological development even though male workers need not necessarily have been conscious of the expendable roles that fragmented social divisions of labor create. (Parallels can be drawn for other forms of division - race, ethnicity and the like). What are the various mechanisms that have been used to exclude women from the technologically more productive and financially more lucrative occupations and job tasks within occupations? Since women have demonstrated time and again the capability to perform a wide range of tasks and to use complex machines, it will be interesting to analyse the various institutional and ideological mechanisms that have permitted the transformation of female into male and male into female specific tasks. The three most important mechanisms were and still are: the level of capitalisation, access to training and continuity/periodicity of work.

While technology increases productivity it requires capital for its adoption. Capital has been used as a mechanism to transfer women from higher productivity and hence from higher wage to lower wage jobs. A few illustrations from studies by labor historians are provided. Bennett (1986) notes that women produced commercial ale - one of the two basic food items in medieval England - after the manufacture of bread, the other basic food item, became professionalised. Baking required ovens and equipment that was relatively expensive and professional bakers were male. It is ironic, however, that females continued to be bakers in the household economy where it remained a labor intensive activity. Ale making remained a lucrative household industry. The equipment required for making ale in the cottage industry phase - large pots, vats and ladle, were available in most households. Moreover, it was a labor intensive activity. The final product spoiled rapidly and transported poorly. It required highly skilled people able to deal with these uncertainties - hence the preponderance of ale-wives. However, when hops were introduced from Europe and beer making became a commercial activity, brewing became man's work and the ideology of woman's nature was used to exclude them. However, women continued to work in the liquor sector as bar maids; i.e., in the service sector rather than in the profitable production sector. This is but one example of such practices that occurred in the textile industry, lace making, garment manufacturing - a whole range of activities that had been considered as commensurate with women's nature and which remained as women's work in the household sphere.

The relegation of women to low technologies was perpetuated through the denial of training to them. Young males were apprenticed and received formal training. Females learned their skills informally from their parents or through domestic service in artisan households. Orphaned children who received formal training in charity institutions learnt crafts that were labor intensive such as spinning or sewing. Boys were taught to read and write - knowledge that was denied to girls - hence the large preponderance of illiterate working women. In printing shops, type setting and pulling the press were male activities since women could not read enough to typeset (Davis, 1986). In the textile factories women were not taught to repair the machines they worked on and their consequent dependence on men was used to justify their lower wages (Bradley, 1986).

How was the relegation to lower skilled jobs institutionalised in society? Ideology was used to justify the exclusion of women from formal training. This related to the definition of femininity and the priority given to household maintenance in favour of labor for the market. A woman's participation in household production of crafts was periodic since her activity was determined by free time left after fulfilling the household tasks that included food preparation, child care and the like. Thus, unwinding the cocoons and preparing the

thread for bobbins was a female task since this could be accomodated with the many household tasks that women had to perform as mistresses, wives and mothers.

Are there any deviations from this? Is there any evidence in labor history of women's competence in male dominated sectors i.e. in cases where women had penetrated into male preserves? Women's production during the craft industrial stage could be of three kinds: they worked as independent artisans and entrepreneurs; as unpaid helpers of male members of the household; and as wage earning artisans. Only in the first case do we find that women had access to capital and technology. In the latter two cases - as household workers and wage labor - they worked at the low technology end of the industry. A brief discussion of women in male preserves can be very illuminating. In Lyon in the sixteenth century, for example, there were highly qualified female entrepreneurs such as printers, publishers, silk manufacturers, barber surgeons and cabinet makers (Davis, 1986). Such independent entrepreneurs were either the daughters or widows of influential and respected crafts families and as such inherited the capital and technology. They were not perceived as threats since there were too few of them and they were maintaining family enterprises for their sons or sons in law. Even so, restrictive actions were taken so as to limit their capacity for growth and development during downturns in their industry. Davis mentions the limitations imposed on female barber surgeons and their journeyman helpers in Lyon during 1540, journeywomen in 1554, and the 1561 and 1583 rulings against female master silk makers.

The textile industry is used to illustrate these concepts in greater detail since textiles and kindred products are industries in which women have been overrepresented. The textile industry absorbed a large percentage of the female labor force in the U.K. when textiles were an important industry in that country. In developing countries today, where much of the textile and apparel industry has migrated, it continues to absorb the majority of women who are employed in manufacturing. Moreover, the textile industry was one of the earliest industries to be affected by technological developments associated with the capitalist production systems.

The textile industry provides a clear example of the shifting sex composition of jobs in both weaving and spinning. Handloom weaving was performed by men when textile production was a cottage industry. Weaving shifted to women during the putting out system associated with proto-industrialization. Handloom weaving once again became a male activity when weaving shifted to factory production. When factories became fully equipped with power looms women, once again, became weavers. Women were spinners in the spinning jenny and water frame stage and men became spinners when the more productive mule spinning was introduced (Hartmann, 1979). Women moved in and out of the tasks of weaving and spinning during different historical periods. This was primarily due to changes in technology.

The specific example of the hosiery industry in Leicester is used to illustrate, in further detail, the switching of tasks between men and women, with the progression of technology and the need to maximise returns to capital. Bradley (1986) and Osterud (1987) provide good documentation of the dynamics of job transformation between female and male workers and the association between the feminisation of specific tasks and technological development. The operation of knitting frames was considered to be a male task in the eighteenth century. However, from the beginning of the nineteenth century there are accounts of women being engaged in operating frames. This was despite the fact that the Charter of 1745 disallowed women (except widows) to work frames. For example, in 1845 the Royal Commission found that 25% of the frames in the town of Leicester, and up to 50% in the surrounding villages, were worked by women. Some individual employers had 66% of their frames operated by women. What underlay the encroachment of women into a male preserve? The total profits that entrepreneurs could make depended partly on the rents they collected on machinery, such as frames. It was profitable for them to distribute their work over a large number of workers. The total work was divided in such a way as to maximise rental incomes from the frames - hence

the importance of female labor. However, in operating these frames women were to be found in the economically unstable and less expensive wrought hose branch while men had begun to shift to knitting in the higher paid glove production.

There was increased feminisation when hosiery production moved to the factory. In the early days of factory production women worked on machines in the new factories. Bradley (1986, p.61) notes that by 1845 women were working in large numbers in the factories operating machines, since men were not willing to accept the regular hours and discipline of the factory. For example, in 1845 Thomas Collins had installed 55 rotary machines and employed a predominantly female workforce to operate these machines. However, the decline of cottage industry due to factory competition brought more males into factory production where they became rotary machine operators by displacing women. Women became relegated to subsidiary and supportive tasks while men became the weavers. By the early 1890's the rotary machines, especially the Cotton's Patents, were almost all worked by men. While 31% of hosiery workers in 1861 and 73.5% by 1901 were female, the majority were now employed in specific female jobs such as seaming, folding and pressing. However the practices of a few firms such as Stretton's, which continued to use women on the Cotton's Patents inspite of union protests, proved that women were competent to use complex machines.

A redefinition of female tasks occurred once again as labor shortages during the 1914-18 war allowed women to work on all types of knitting machines and enter into other higher paid male preserves such as countering and trimming. Women formed the major part of the workforce in the hosiery industry until the mid twentieth century. However, men once again became the knitters as improvements in knitting technology increased productivity and wages. In 1983, when a single person was able to tend 18 machines rather than one machine, and wages increased in that sector, men again became the knitters in the sock industry in Leicester. In summary, knitting was a woman's task during the cottage industry phase when frames were rented by the household; they worked in factories when the narrow frame had become obsolescent with the development of the more efficient wide frame which men began to operate; they also operated wide frames during the world wars when there was a shortage of male labor. Women were used on new technology when men refused and were supplanted when its inevitability was demonstrated. They were displaced to old technology so that obsolescent technology could be retired at a rate which could still provide profits to entrepreneurs.

What is true of textiles is true of other industries as well. Abbott (1969) mentions that when cigar making was a home industry in the U.S. before 1800, it was a woman's activity in the agricultural household. Early factories employed women who were later displaced by male immigrants willing to work for similar low wages. By 1860 only 9% of the workers were women in an industry that had previously been dominated by women. In 1869, when the wooden mould was introduced from Europe, women became the primary workers. Similar gender switchings, related to technology, have been documented for the printing industry (Baker, 1964; Hartmann, 1979). This theme of the association between technology, job switching and women's employment will continue in the next section where it shall be documented for the service sector.

The well specified female and male areas of activity that characterise gender concentrations in occupations at any point in time is primarily related to the needs of capital. Technological development is a manifestation of this need. Sexual stereotyping and sex segregation in occupations are institutionally created and ideologically supported. Occupational sex segregation and ideology of the nature of women have been used to exclude women from new and productive technologies. Patriarchial values have been the instrumental variables.

4. THE ROLE OF WOMEN IN THE DEVELOPMENT OF TECHNOLOGY IN THE SERVICE SECTORS

The steady growth in the participation of women in the contemporary labour force is related to sectoral shifts and the increasing importance of the service sector in the economy. Women moved into the service sector with the progressive mechanization of the female dominated goods producing sectors such as textiles and the decline of the family farm. For example, in the U.S. female non agricultural employment in domestic and personal services declined from 67% in 1870 to 32% in 1930.

Employment in the service sector is also characterized by sex segregation and low wages. Feminization of a particular service is also directly related to technological developments in that particular service. The introduction of improved processes and related employment displacements is also associated with the progressive development of technology in the service sector. The clerical sector is used to illustrate the relationship between technology and women's employment in the service sector and the arguments made in the paper thus far are continued.

The 1910's were characterized by a significant occupational redistribution with a massive incorporation of women into service occupations. In 1870 women accounted for 0.2% of the labor force in the clerical occupations. By 1980, four out of five clerical workers were women (Maillier and Rosser, 1987). In 1911 clerical work provided jobs to 3% of the total female labor force; by 1981 more than 33% of a much larger labor force was employed in that sector. As in manufacturing, technology - in the form of new machines - went hand in hand with new ways of organizing work. Feminization of the clerical sector was as intrinsically related to machines as it was to new patterns in the organization of work.

What is the relation between technical advances and the increased participation of women in clerical occupations? The feminization of the clerical work force occurred in the late nineteenth and early twentieth centuries when the character of clerical work and its organizational structure changed radically (Davies, 1982). The traditional office was small with a low level of technology, characterized by close working relations between clerks and employers. In the U.S. in the pre-civil war period the office was a male preserve and there were no women workers (Davies, p. 27). The typical clerk, prior to this transformation, was an apprentice manager or businessman. He performed a wide variety of tasks in order to learn different facets of the business. Clerical work was viewed as temporary and a period of preparation for higher and more rewarding tasks. In the late nineteenth and early 20th centuries clerical work underwent dramatic changes in the type of technology used by workers and the organization of the workplace. Technologies such as the typewriter, telegraph, telephone, calculating and duplicating machines revolutionized the office. This resulted in the need for a larger workforce, standardization of tasks and greater specialisation among workers. By the end of the transformation clerical work was primarily a dead end occupation; most clerical workers would remain in their positions during their working lives. This transformation was accompanied by an enormous expansion of clerical work, changing sexual composition of the workforce and the introduction of new technology that made clerical work into a series of routine, repetitive tasks and female clerical workers into office operatives. In the clerical sector the development of typewriters and telephones had much to do with the feminization of the work force, since the typewriter vastly increased the demand for correspondence and the keeping of records i.e. filing. Feminization proceeded at different rates within the clerical field. It proceeded most rapidly amongst typists and stenographers - both occupations were marked by new technological developments. In 1880 women constituted 40% of all stenographers and typists. In 1900 and 1930 the proportion had increased to 75% and 95% respectively. Bookkeepers and accountants continued to be male and it is only in the last two decades that there has been a significant change in these clerical tasks.

Machines did not cause the routinization of work, though they certainly facilitated it. The successful invention and manufacture of the typewriter resulted from developments in the organization of work that made these machines necessary. Women facilitated their rapid diffusion because they adapted to the conditions of work and accepted the rigid hierarchial structure of authority. The rigid hierarchial structures within an office diminished opportunities for upward mobility. A typist who was involved in a restricted and highly specialised task lost the opportunity to gain knowledge that would allow her to take decisions over a range of tasks that the management role implied. Feminization of the clerical work facilitated technological change.

Female employment is likely to be significantly influenced by new technological developments such as microelectronics. New developments in information processing will cause the elimination of much of the clerical staff, particularly those involved in recording, storing and retrieval of information (Maillier and Rosser, 1987).

The earlier round of technology increased job opportunities in a variety of new sectors, either through the replacement of males or in new tasks opened up by technology i.e. typing, or telephone switch board operating. The elimination of paper as a medium of information storage and the storage of information on disks will cause a basic restructuring of the work force and severe cutbacks among typists and file clerks. Higher levels of training in computers and word processors are likely to lead to a demand for a higher paid and more male work force. While this is not a negative trend, and a better sexual balance in all jobs is desirable, most of the jobs that are eliminated will belong to women typists and clerks.

5. CONCLUSION

Society has witnessed a deluge of technological developments since the dawn of the Industrial Revolution. On the aggregate these developments, relative to earlier societies, have resulted in increased employment opportunities, increased productivity, increased prosperity and higher levels of material and non material consumption for both men and women. The increase in the paid labor force participation of women is a direct consequence of the increasing incorporation of improved technologies in production processes and the improved organization of work that this often entailed. However, two factors have been constant in this incorporation. First, due to the patriarchial form of societal structure, women have been assigned to the lower technology end of production processes. As new innovations were adopted women were reassigned to tasks which had relatively more obsolete technologies. Second, they have been paid lower wages than men for their effort. As the aggregate increase in employment has been accompanied by sectoral declines, women have been concentrated in the declining sectors due to their assignment to the low technology end of production. This paper argues that women have facilitated the adoption of new technology through their positive adjustments to displacements and "acceptance" of lower levels of technology and lower wages.

There is nothing that is deterministic about the relation between women and technology. The various arguments that have been put forward in favor of women's assignment to low technology jobs, such as their weaker physical constitution, poorer mechanical skills, inability to handle complex tasks, loss of femininity and other similar sexist justifications do not stand the test of historical analysis. Even a cursory glance at the relation between women and technology from a historical perspective provides evidence that there has been no task that has been too complex for women or no technology too onerous for them to use when appropriate training has been provided. Occupational segregation with its attendant discrimination in training opportunities, access to improved technologies, and lack of comparable wages has perpetuated itself through relations of dominance and subdominance based on gender, race, region of origin and like variables. The key question is whether greater awareness of women's issues will make

this period of rapid technological change significantly different from earlier periods of technological progress.

Past advances in technology have decreased labor demand in some areas and increased them in others. We can expect similiar dislocations in the new round of technological developments that are currently underway. For example, in the manufacturing sector computer controlled technology will tend to eliminate repetitive jobs. The effect of this will be felt disproportionately by women since they tend to be concentrated in jobs that involve repetitive tasks and that have a high potential for automation. Maillier and Rosser (1987, p.159) note that the repetitive assembly of electronic goods in female dominated industries is likely to be affected by new technology. In the dawn of the new information age and its related microtechnologies there is an opportunity to eradicate the unequal and socially unjust use of women's labor which has been discussed in this paper. The advent of these new technologies has the potential for raising the quality of life for those in the manufacturing and service sectors. However, there is also a great potential for increasing social inequalities in relation to both gender and class. The outcome will depend on those who have access to the opportunities that open up as a result of the new technological developments.

REFERENCES

Abbott, E., 1969, *Women in Industry*, Arno Press, N.Y.

Baker, E., 1964, *Technology and Woman's Work*, Columbia University Press, N.Y.

Bennett, J. M., 1986, "The Village Ale-Wife: Women and Brewing in Fourteenth-Century England" in B. Hanawalt, (ed.), *Women and Work in Preindustrial Europe*, Indiana University Press, Bloomington, U.S.A.

Berg, M., P. Hudson and M. Sorenschor, 1983, *Manufacture in Town and Country Before the Factory*, Chapter 1, Cambridge University Press, New York.

Berg, M., 1985, *The Age of Manufactures 1700-1820*, Chapters 3, 6 and 8, Fontana Press, London.

Bergman, B., 1986, *The Economic Emergence of Women*, Basic Books, New York.

Bradley, H., 1986, "Technological Change, Management Strategies, and the Development of Gender-based Job Segregation in the Labor Process", in D.Knight and H. Willmott, (eds.), *Gender and the Labor Process*, Gower, U.K.

Chatterjee, L. and T.R. Lakshmanan, 1987, "Facilitating Technical Change: The Role of Women" in M. Fischer and P. Nijkamp, (eds.), *Regional Labour Markets*, North Holland, Amsterdam.

Dauber, R. and M. Cain, (eds.), *Women and Technological Change in Developing Countries*, Westview Press, Boulder, Colorado.

Davies, M., 1982, *Woman's Place is at the Typewriter: Office Work and Office Workers 1870-1930*, Temple University Press, Philadelphia.

Davis, Z., 1986, "Women in the Crafts in Sixteenth-Century Lyon" in B. Hanawalt, (ed.), *Women and Work in Preindustrial Europe*, Indiana University Press, Bloomington, U.S.A.

Feldberg, R., 1983, "Comparable Worth: Toward Therory and Practice in the U.S.", in B. Gelpi, N. Hartsock, C. Novak and M. Strober, (eds.), *Women and Poverty*, University of Chicago Press, Michigan.

Glazer, N., (1980), "Everyone Needs Three Hands: Doing Unpaid and Paid Work" in S.Berk, (ed.), *Women and Household Labor*, Sage Pub., California.

Hanawalt, B., (ed.), *Women and Work in Preindustrial Europe*, Indiana University Press, Bloomington, USA.

Hartmann, H., 1979, "Capitalism, Patriarchy and Job Segregation" in Z. Eisenstein, (ed.), *Capitalist Patriarchy and the Case for Socialist Feminism*, Monthly Review Press, New York.

Humphrey, J., 1985, "Gender, Pay and Skill: Manual Workers in Brazilian Industry" in H. Afshar, (ed.), *Women, Work and Ideology in the Third World*, Chapter 9, Tavistock, New York,

I.L.O. Office of Women, 1981,, "Women, Technology and the Development Process" in R. Dauber and M. Cain, (eds.), *Women and Technological Change in Developing Countries*, Westview Press Boulder, Colorado, pp. 33-47.

Joekes, S., 1985, "Working for Lipstick? Male and Female Labor in the Clothing Industry in Morocco" in H. Afshar, (ed.), *Women, Work and Ideology in the Third World*, Chapter 8, Tavistock, New York.

Joshi, H. and V. Joshi, 1976, *Surplus Labor and the City: A Study of Bombay*, Oxford University Press, Delhi.

Knight, D. and H. Willmott, (eds.), *Gender and the Labour Process*, Gower, UK.

Maillier, A.T. and M.J. Rosser, 1987, *Women and the Economy*, St. Martin's Press, New York.

Mann. J. and B. Hellwig, 1988, "The Truth About Salary Gap(s)" in *Working Women*, January, pp. 61-62.

Mather, C., 1985, "Rather than Make Trouble, It's Better Just to Leave: The Lack of Industrial Strife in West Java" in H. Afshard, (ed.), *Women, Work and Ideology in the Third World*, Chapter 7, Tavistock Publications, New York.

Matthaei, J., 1982, *An Economic History of Women in America: Women's Work, the Sexual Division of Labor and the Development of Capitalism*, Schocken Books, New York.

OECD, 1984 Employment Outlook.

Osterud, N., 1986,"Gender Divisions and the Organization of Work in the Leicester Hosiery Industry" in A.V. John, (ed.), *Unequal Opportunities: Women's Employment in England 1800-1918*, Basil Blackwell.

Reskin, B., 1986, *Sex Segregation in the Workplace*, National Academy Press, Washington D.C.

Reskin, B. and H. Hartmann, (eds.), 1985, *Women's Work, Men's Work: Sex Segregation on the Job*, National Academy Press, Washington D.C.

Roos, P. and B. Reskin, 1984, "Institutional Factors Contributing to Sex Segregation in the Workplace" in B. Reskin, (ed.), *Sex Segregation in the Workplace*, Chapter 13, National Academy Press, Washington D.C.

Treiman, D. and H. Hartmann, (eds.), 1981, *Women, Work and Wages: Equal Pay for Jobs of Equal Value*, National Academy Press, Washington D.C.

UNIDO, 1980, *Women in the Redeployment of Manufacturing Industry to Developing Countries*, Paper on Structural Change 18.

Urquhart, M., 1984, "The Employment Shift of Services: Where did it Come From?", *Monthly Labor Review*, April, pp.15-22.

Advances in Spatial Theory and Dynamics
A.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 6

An Oil-Exporting Region versus an Industrialized Region

J.M. Hartwick and M. Spencer

1. INTRODUCTION

Can an oil importing region or country benefit from an exogenous increase in the price of the oil it imports? The simple answer is yes. That is, if by paying more for oil imports, it increases the income of the oil exporter enough so that the oil exporter rapidly expands *its imports* from the oil importing country. More informally, OPEC could raise the price of its oil exports, receive more income, and then expand its imports of, say, industrial goods by more than the increase in its oil revenues. The oil importer's economy can gain more revenue than it loses through increased payments for oil imports. Chichilnisky (1983) investigated this phenomenon in a two region or country model. In Hartwick (1984) the sensitivity of her numerical results was investigated and the sensitivity of her results to her choice of numeraire price was analyzed. In a two-region model the price of industrial goods is the same in both regions so that price was selected as the numeraire. In this case, the oil importing region never benefited from higher oil prices. In the jargon of Chichilnisky, the "substitution effect" disappeared. But the basic Chichilnisky model seemed unsatisfactory because commodity balance was absent in the industrial goods sector and the link between capital goods and industrial goods (acting as investment) was inadequate. We shall redevelop the basic model and exhibit the "substitution effect" in the revised model.

There is a Keynesian element in the substitution effect besides the terms of trade or relative price effect noted above. Unemployed capital goods are brought into production as oil prices rise. Also we note that a rising GNP or gross domestic product is not a good index of rising welfare. In the model below both wages and employment rates decline as the exogenous price for oil imports rises, so workers fare poorly in the face of rising oil prices. The model is special: the oil importing country's economy is fully specified, but the economy of the oil exporting country is specified only to the extent that it sets the price of its oil exports and imports industrial goods at an endogenously determined price. There are two sectors in the oil importing country - an industrial goods and a basic or consumer goods sector. Supply schedules with less than infinite price elasticities for labor and capital for the oil importing country are given. The model is stationary given these factor supply characteristics: capital is not accumulated nor does the population grow. In the Appendix we set out the Chichilnisky model and briefly note the differences between that model and the one that we describe immediately below.

2. THE MODEL

The two commodities produced in the oil importing region have Leontief production functions.

$$B = \min (L^B/a_1, O^B/b_1, K^B/c_1) \tag{1}$$

where L^B, O^B, and K^B denote quantities of labor, oil, and capital used in the production of the consumption good in quantity B. a_1, b_1, and c_1 are technical factor productivity coefficients. For the industrial good, we have

$$I = \min (L^I/a_2, O^I/b_2, K^I/c_2) \tag{2}$$

where definitions are analogous to those for the consumption goods production function.

The associated 'dual' competitive price equations are

$$p_B = a_1 w + b_1 p_O + c_1 r p_I \tag{3}$$

and

$$p_I = a_2 w + b_2 p_O + c_2 r p_I \tag{4}$$

where w is the wage rate, p_O the price of oil, r the rate of return for reproducible capital or the interest rate, p_I the price of industrial (new capital) goods and p_B the price of the consumption goods. $p_I r$ is a proxy for the user cost or the rental price of capital.

In spite of the fixity of production coefficients, the relative prices of p_B and p_I will change as p_O changes resulting in substitutability of industrial and consumer goods.

Since all wage income will be assumed spent on consumer goods, the labor supplied will respond to its wage expressed in terms of its purchasing power.

$$L = \alpha \cdot (w/p_B), \quad \alpha > 0 \tag{5}$$

and existing capital utilized depends on the rate of return r, i.e., $K = \beta r$. Total capital utilized comes from the existing stock of say different vintages plus new capital goods I^D covering the depreciation on existing capital goods in use. Thus

$$K^S = \beta r + I^D \quad \beta > 0. \tag{6}$$

Demand behavior is such that all wage income is spent on consumer goods

$$p_B B^D = wL \tag{7}$$

where the superscript D denotes *domestic* consumption by the oil importing region. Labor, oil and capital demands are respectively

$$L^D = a_1 B^S + a_2 I^S \tag{8}$$

$$O^D = b_1 B^S + a_2 I^S \tag{9}$$

$$K^D = c_1 B^S + c_2 I^S. \tag{10}$$

Domestic demand for industrial goods produced derives from depreciation on capital goods currently utilized. Thus

$$I^D = \gamma K^D. \tag{11}$$

Implicit here is the assumption that as existing capital goods are being utilized, they depreciate instantaneously at an exogenous rate γ. This depreciated amount is replaced by new capital goods from the current production of industrial goods. A steady state comes to mind. (A different assumption would be depreciation equal to $\gamma r\beta$, or only "old" capital goods currently depreciate in use.)

Material balance requires that

$$K^D = K^S \tag{12}$$

$$L^D = L^S \tag{13}$$

$$B^D = B^S \tag{14}$$

$$O^D = O^S \tag{15}$$

$$X = I^S - I^D. \tag{16}$$

The last two equations indicate oil used by the importing nation equals oil supplied from abroad (the other country) and industrial goods currently produced and not used by the oil importer, are exported to the oil supplier.

Equations (3) to (16) form our two country model. The price of basic goods is our numeraire or $p_B = 1$. We have fourteen unknowns: B^S, B^D, I^S, I^D, L^S, L^D, K^S, K^D, O^S, O^D, X, w, r, p_I. This is the model we solved, using parameter values from Chichilnisky (1983). Before taking up our results (computer outputs) we report that whether the value of exported industrial goods equals the value of imported oil depends on the value of the rate of depreciation (a parameter to be selected) relative to the rate of interest (a variable solved in the model). We then have

Proposition: The value of oil imports exceeds (equals, fall short of) the value of industrial goods exported as the rate of interest falls short of (equals, exceeds) the rate of depreciation.

Proof: Using (15) and (9) we have

$$P_O O^S = b_1 B^S P_O + b_2 I^S P_O = B^S(P_B - a_1 w - c_1 r P_I) + I^S(P_I(1 - c_2 r) - a_2 w)$$

using (3) and (4)

$$= wL^D - w[L^D - a_2 I^S] + rP_I[K^D - c_2 I^S] + P_I I^S - c_2 I^S r P_I - a_2 w I^S$$

using (7), (14), (8) and (10)

$$= P_I(I^S - rK^D)$$

$$\gtrless P_I(I^S - \gamma K^D) \quad \text{as } r \gtrless \gamma$$

where the value of industrial goods exported is $P_I(I^S - I^D)$ and $I^D = \gamma K^D$ from (11).

The treatment of rental income from ownership of the employed capital stock is obviously special. Wage income "soaks up" the value of basic goods produced leaving rental income from capital to purchase current industrial goods produced net of oil imports. (This follows from the fact that our Z (= $wL + rp_I I^S$) equals GNP (= $p_B B^S + p_I I^S - p_O O^S$)). In the special case of current account balance rental income from capital

equals the value of industrial goods used domestically. If one supposes that capital is owned by the public sector, the treatment seems most reasonable. The main point is that there is no current consumption out of income from capital by a rentier class. For the case of a current account surplus, one might hypothesize that the income corresponding to the surplus is being directed by a rentier class to consumption of foreign commodities but this link is not explicit in the model.

We are interested in exhibiting a certain mode of behavior of the model - to display a logical possibility. We make use of Chichilnisky's parameters which are constrained in part as follows:

(i) $M = a_1b_1 - a_2b_1 > 0$ the consumption or basic sector is relatively labor intensive compared with the oil intensiveness of the industrial sector.
(ii) c_1 is small or the consumption good requires very few capital inputs. (By setting $c_1 = 0$, Chichilnisky obtains comparative static results analytically. For $c_1 > 0$, she appeals to numerical examples as we do.)
(iii) b_1 is small or the consumption or basic good requires little oil in production.

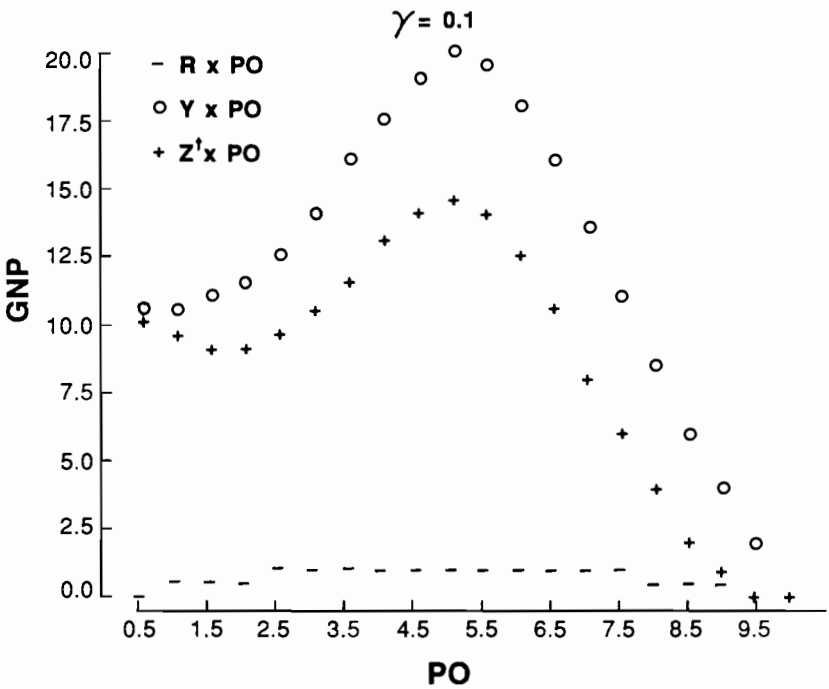
These assumptions imply that the direct link between the price of oil and the price of the basic good is very weak. In fact, since $p_B = 1$ (numeraire assumption) an increase in the price of oil will directly lower the wage rate and given Assumption (iii) this effect will be relatively small. For c_1 insufficiently small, no "substitution effect" is observed (see Hartwick (1984)).

3. THE RESULTS

In this model, we have introduced a new technical coefficient, the rate of depreciation. We indicated that the value of this parameter directly affected the trade deficit of the oil importing country. It turns out that for small rates of depreciation (the oil importing country with a current account trade surplus) rising oil prices result at first in Chichilnisky's "substitution effect" and then in the "income effect". In other words, our model displays all of the characteristics of Chichilnisky's model, which as we have already noted, is flawed. For larger values of the rate of depreciation no "substitution effect" of rising oil prices is observed.

The three runs[1] we described here make use of the parameters set out in Chichilnisky (1983; Appendix A4). The rate of depreciation, the parameter peculiar to our version of the two country model, is set at $\gamma = 0.1$, 0.4 and 0.8 in three runs. When $\gamma = 0.1$ the oil importer always has a current account surplus and both GNP and GDP (this latter being the value of all outputs produced by the oil importing country: GNP is GDP net of imports (here of oil) and of course equals the value of domestic factors of production in use - namely $wL + rp_IK$, which we indicate by Z.) In brief, for $\gamma = 0.1$, we see GNP, GDP, and Z eventually rise over the "early" range of exogenously increasing oil prices. The Chichilnisky "substitution effect" is exhibited. See Table 1 and Figure 1. For $\gamma = 0.4$, the rate of interest, which varies, is on average slightly higher than γ and only GDP rises in the "early" range of increasing oil prices. See Table 2 and Figure 2. At first there is a small current account surplus for the oil importer, then a deficit and then again a surplus. For $\gamma = 0.8$, GNP, Z and GDP decline as oil prices rise. No substitution effect is exhibited. See Table 3 and Figure 3.

[1]These computer runs were solved as a full fourteen nonlinear equation system with the code ZSCNT of the large package IMSL. The parameters used are $a_1 = 0.2$, $a_2 = 0.2$, $b_1 = 0.1$, $b_2 = 0.2$, $c_1 = 0.001$, $c_2 = 0.6$, $\alpha = 1.0$ and $\beta = 2.0$, $p_B = 1.0$. The depreciation rate γ assumes the values 0.1, 0.4, 0.8 in three separate runs. The price of oil p_0 moves in each run from 0 to 10 as in Chichilnisky.



† Since GNP = Z, the GNP series coincides with the + series and the computer does not print both.

Figure 1 Low Depreciation Rate and the Price of Oil Rises

The oil importing country always runs a current account deficit.

Why does the run with $\gamma = 0.1$ exhibit the substitution effect and the one with $\gamma = 0.8$ not? Consider relative prices as the price of oil changes. For $\gamma = 0.1$ the ratios rP_I/w and P_I/p_O rise in the "early" range of oil price increases and then decline. See Table 1. (P_I/p_O is indicated by TOT for Terms of Trade in the Tables.) For $\gamma = 0.8$ these ratios only move in one direction as p_O rises. See Table 3. This same monotonicity in relative prices holds for $\gamma = 0.4$. See Table 2. Clearly then the magnitudes of as well as signs of changes in relative prices matter. Intuitively one thinks of the "substitution effect" occurring because the terms of trade, p_I/p_O, move strongly in favor of the oil importing country and then move against this country in the "income effect" range of rising oil prices. This intuition is borne out in the data. Common to all three cases is a rise and then decline in r, rP_I, I^S, I^D, X, $K^S(=K^D)$, and the value of exports, p_IX. They all peak at about the same value of the price of oil. We ask the reader to verify these statements by examining Tables 1,2, and 3 corresponding to runs with $\gamma = 0.1$, 04, and 0.8 respectively.

Is the "substitution effect" logically connected to a current account surplus in the sense that the "substitution effect" will not be observed without a current account surplus? Of course we ask this question within the confines of this model. We do not know. For $\gamma = 0.4$, GDP rose as a current account surplus was observed but both GNP and

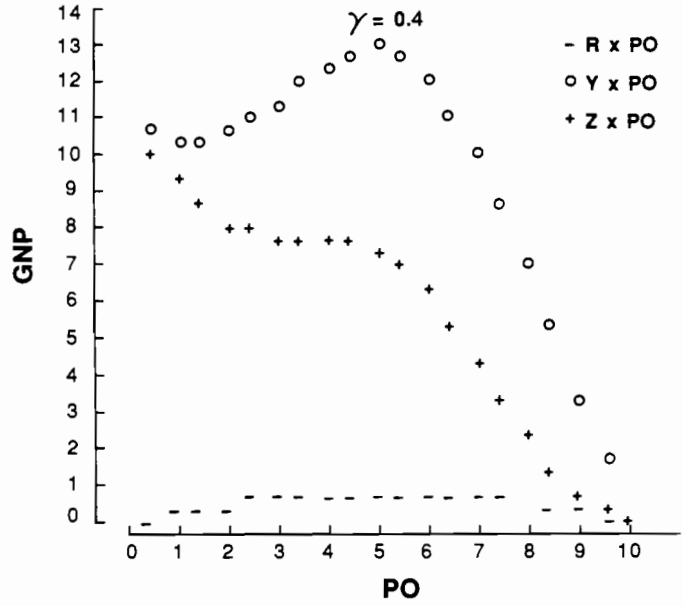


Figure 2 Medium Depreciation Rate and the Price of Oil Rises

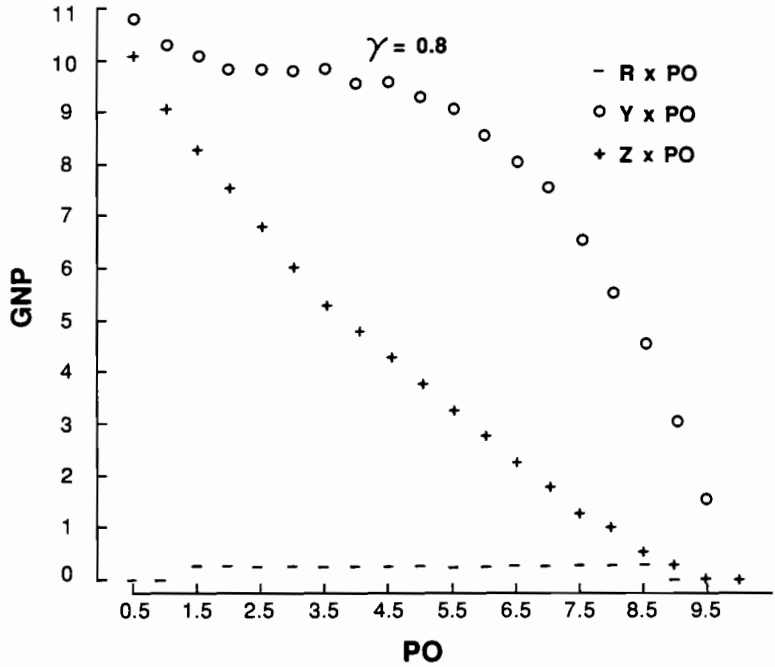


Figure 3 High Depreciation Rate and the Price of Oil Rises

national income (Z equal to GNP) declined over this range of oil price increases. The surplus turned to a deficit and back to a surplus over the range of oil price increases. One might argue that extra industrial goods exports has a multiplier effect on aggregate supply and thus the trade surplus is necessary for GNP to rise as oil import prices rise.[2] Certainly both the value of oil imports and current surplus increase over the phase in which the "substitution effect" is exhibited. They decline over the phase in which the income effect is exhibited. Of course, as we have noted, numerous series rise and fall in parallel with GNP so one should perhaps not pinpoint a single source of the "substitution effect". Both relative price effects and real effects mingle. We are dealing with a small general equilibrium system and pinpointing causal links is difficult.

4. CONCLUDING REMARKS

Rapidly rising oil prices caused many economists to reflect on the effects these increases could have on oil importing countries. The "Dutch Disease" (or "Gregory syndrome") was the adverse effects *on oil exporters* of a rising relative price of the oil exporters currency. The rising value of the currency led to a shrinking of international demand for industrial exports from the oil exporting country. See for example Corden and Neary (1982) and Corden (1983). Chichilnisky alerted economists to the logical possibility of non-deleterious effects of oil price imports on an oil importing nation. She coined the term "substitution effect" for an instance of rising prices of oil imports and a rising GNP in the oil importing country. In this paper we have exhibited these "substitution effects" in a restructured Chichilnisky model. The logical basis of such effects has been placed on a sounder footing.

[2] This argument, that a disequilibrium must exist for the "substitution effect" to manifest itself, carries over to Chichilnisky's model where a material imbalance exists in the industrial good sector. See the Appendix below and Hartwick (1984). Above, this disequilibrium is in the trade balance. Chichilnisky's material imbalance (the size of her I^D) increases over the phase of rising oil prices over which the "substitution effect" is exhibited and declines over the phase over which the "income effect" is exhibited. Our current account imbalance moves in "parallel" with her industrial goods material imbalance.

Table 1 Low Depreciation Rate and the Price of Oil Rises

A1 = 0.30 A2 = 0.20 B1 = 0.10 B2 = 0.20 C1 = 0.001 C2 = 0.60
ALPHA= 1.00 BETA = 2.00 PB = 1.00 GAMMA = 0.10

PO	PI	W	R	RPI	RPI/W
1.0000	1.0611	2.9985	0.4106	0.4357	0.1453
2.0000	1.6530	2.6627	0.7264	1.2008	0.4510
3.0000	2.4830	2.3254	0.9517	2.3630	1.0162
4.0000	3.4316	1.9876	1.0850	3.7234	1.8733
5.0000	4.1017	1.6513	1.1262	4.6192	2.7974
6.0000	4.1341	1.3185	1.0769	4.4521	3.3766
7.0000	3.6604	0.9885	0.9392	3.4379	3.4777
8.0000	3.0309	0.6595	0.7143	2.1651	3.2832
9.0000	2.4582	0.3300	0.4015	0.9869	2.9903
10.0000	2.0000	0.0000	0.0000	0.0000	0.0000

PO	IS	ID	X	OD	BS
1.0000	1.5058	0.0912	1.4146	1.2003	8.9913
2.0000	2.6786	0.1614	2.5172	1.2447	7.0898
3.0000	3.5157	0.2115	3.3042	1.2439	5.4077
4.0000	4.0121	0.2411	3.7710	1.1975	3.9505
5.0000	4.1664	0.2503	3.9162	1.1059	2.7266
6.0000	3.9857	0.2393	3.7464	0.9709	1.7379
7.0000	3.4769	0.2087	3.2682	0.7931	0.9772
8.0000	2.6449	0.1587	2.4862	0.5725	0.4349
9.0000	1.4868	0.0892	1.3976	0.3083	0.1089
10.0000	0.0000	0.0000	0.0000	0.0000	0.0000

PO	LD	KD	Z	PIX	TOT*
1.0000	2.9985	0.9125	9.3888	1.5010	1.0611
2.0000	2.6627	1.6143	9.0281	4.1609	0.8265
3.0000	2.3254	2.1148	10.4051	8.2044	0.8277
4.0000	1.9876	2.4112	12.9284	12.9404	0.8579
5.0000	1.6513	2.5026	14.2865	16.0630	0.8203
6.0000	1.3185	2.3932	12.3930	15.4878	0.6890
7.0000	0.9885	2.0871	8.1524	11.9629	0.5229
8.0000	0.6595	1.5874	3.8717	7.5355	0.3789
9.0000	0.3300	0.8922	0.9895	3.4356	0.2731
10.0000	0.0000	0.0000	0.0000	0.0000	0.2000

PO	TRADE SURPLUS	Y	GNP
1.0000	0.3007	10.5891	9.3889
2.0000	1.6715	11.5175	9.0281
3.0000	4.4727	14.1373	10.4055
4.0000	8.1505	17.7184	12.9285
5.0000	10.5333	19.8161	14.2864
6.0000	9.6623	18.2150	12.3895
7.0000	6.4112	13.7041	8.1524
8.0000	2.9557	8.4515	3.8717
9.0000	0.6612	3.7638	0.9895
10.0000	0.0000	0.0000	0.0000

* Terms of trade or P_I/p_O

Table 2 Medium Depreciation Rate and the Price of Oil Rises

A1 = 0.30 A2 = 0.20 B1 = 0.10 B2 = 0.20 C1 = 0.001 C2 = 0.60
ALPHA = 1.00 BETA = 2.00 PB = 1.00 GAMMA = 0.40

PO	PI	W	R	RPI	RPI/W
1.0000	0.9566	2.9991	0.2733	0.2615	0.0872
2.0000	1.3139	2.6645	0.4833	1.6350	0.2383
3.0000	1.7187	2.3297	0.6329	1.0878	0.4669
4.0000	2.1155	1.9949	0.7221	1.5276	0.7658
5.0000	2.4242	1.6606	0.7508	1.8201	1.0961
6.0000	2.5783	1.3271	0.7194	1.8549	1.3977
7.0000	2.5665	0.9946	0.6284	1.6127	1.6214
8.0000	2.4293	0.6628	0.4780	1.1613	1.7521
9.0000	2.2248	0.3313	0.2686	0.5976	1.8035
10.0000	2.0000	0.0000	0.0000	0.0000	0.0000

PO	IS	ID	X	OD	BS
1.0000	1.5035	0.3644	1.1390	1.2002	8.9948
2.0000	2.6731	0.6444	2.0287	1.2446	7.0998
3.0000	3.5072	0.8439	2.6633	1.2442	5.4275
4.0000	4.0050	0.9628	3.0422	1.1990	3.9797
5.0000	4.1666	1.0011	3.1655	1.1091	2.7576
6.0000	3.9939	0.9592	3.0346	0.9749	1.7613
7.0000	3.4892	0.8378	2.6514	0.7968	0.9893
8.0000	2.6550	0.6374	2.0176	0.5749	0.4393
9.0000	1.4920	0.3581	1.1339	0.3094	0.1098
10.0000	0.0000	0.0000	0.0000	0.0000	0.0000

PO	LD	KD	Z	PIX	TOT
1.0000	2.9991	0.9111	9.2329	1.0896	0.9566
2.0000	2.6645	1.6109	8.1227	2.6654	0.6569
3.0000	2.3297	2.1098	7.7226	4.5774	0.5729
4.0000	1.9949	2.4070	7.6566	6.4360	0.5289
5.0000	1.6606	2.5027	7.3128	7.6738	0.4848
6.0000	1.3271	2.3981	6.2095	7.8243	0.4297
7.0000	0.9946	2.0945	4.3671	6.8049	0.3666
8.0000	0.6628	1.5934	2.2898	4.9015	0.3037
9.0000	0.3313	0.8953	0.6448	2.5227	0.2472
10.0000	0.0000	0.0000	0.0000	0.0000	0.2000

PO	TRADE SURPLUS	Y	GNP
1.0000	-0.1105	10.4330	9.2329
2.0000	0.1762	10.6118	8.1227
3.0000	0.8448	11.4554	7.7228
4.0000	1.6401	12.4525	7.6566
5.0000	2.1284	12.8582	7.3129
6.0000	1.9749	12.0588	6.2094
7.0000	1.2276	9.9445	4.3671
8.0000	0.3020	6.8892	2.2897
9.0000	-0.2618	3.4293	0.6448
10.0000	0.0000	0.0000	0.0000

Table 3 High Depreciation Rate and the Price of Oil Rises

A1 = 0.30 A2 = 0.20 B1 = 0.10 B2 = 0.20 C1 = 0.001 C2 = 0.60
ALPHA = 1.00 BETA = 2.00 PB = 1.00 GAMMA = 0.80

PO	PI	W	R	RPI	RPI/W
1.0000	0.8455	2.9997	0.0909	0.0769	0.0256
2.0000	1.0329	2.6661	0.1608	0.1661	0.0623
3.0000	1.2208	2.3325	0.2106	0.2572	0.1102
4.0000	1.4021	1.9989	0.2405	0.3372	0.1687
5.0000	1.5686	1.6654	0.2503	0.3926	0.2357
6.0000	1.7132	1.3320	0.2401	0.4113	0.3088
7.0000	1.8303	0.9987	0.2099	0.3843	0.3848
8.0000	1.9170	0.6656	0.1599	0.3065	0.4604
9.0000	1.9729	0.3327	0.0899	0.1773	0.5329
10.0000	2.0000	0.0000	0.0000	0.0000	0.0000

PO	IS	ID	X	OD	BS
1.0000	1.5004	0.7274	0.7730	1.2000	8.9989
2.0000	2.6683	1.2865	1.3818	1.2445	7.1082
3.0000	3.5017	1.6852	1.8165	1.2444	5.4404
4.0000	4.0011	1.9237	2.0774	1.1998	3.9955
5.0000	4.1667	2.0022	2.1644	1.1107	2.7734
6.0000	3.9987	1.9208	2.0779	0.9772	1.7741
7.0000	3.4974	1.6796	1.8179	0.7992	0.9974
8.0000	2.6636	1.2789	1.3847	0.5770	0.4431
9.0000	1.4976	0.7190	0.7787	0.3106	0.1107
10.0000	0.0000	0.0000	0.0000	0.0000	0.0000

PO	LD	KD	Z	PIX	TOT
1.0000	2.9997	0.9092	9.0684	0.6536	0.8455
2.0000	2.6661	1.6081	7.3752	1.4273	0.5164
3.0000	2.3325	2.1065	5.9821	2.2176	0.4069
4.0000	1.9989	2.4047	4.8062	2.9126	0.3505
5.0000	1.6654	2.5028	3.7560	3.3952	0.3137
6.0000	1.3320	2.4010	2.7618	3.5599	0.2855
7.0000	0.9987	2.0995	1.8042	3.3272	0.2615
8.0000	0.6656	1.5986	0.9330	2.6545	0.2396
9.0000	0.3327	0.8987	0.2701	1.5363	0.2192
10.0000	0.0000	0.0000	0.0000	0.0000	0.2000

PO	TRADE SURPLUS	Y	GNP
1.0000	-0.5464	10.2675	9.0676
2.0000	-1.0617	9.8642	7.3753
3.0000	-1.5156	9.7153	5.9821
4.0000	-1.8865	9.6053	4.8062
5.0000	-2.1582	9.3093	3.7560
6.0000	-2.3030	8.6247	2.7618
7.0000	-2.2674	7.3988	1.8042
8.0000	-1.9617	5.5492	0.9329
9.0000	-1.2591	3.0655	0.2701
10.0000	0.0000	0.0000	0.0000

Appendix: The Chichilnisky Model

To work back from the model we have described in this paper to the model which first presented Chichilnisky's substitution effect (rising GNP to the oil importer as oil prices rise), one substitutes for our (6) the following capital supply equation:

$$K^S = \beta r \qquad \beta > 0. \tag{6'}$$

In place of our equation (11) for domestic demand for industrial goods, one has

$$p_O O^S = p_I X \tag{11'}$$

or the value of oil imports equals the value of industrial goods exported. Material balance does not obtain in this version because there is no domestic demand specified for industrial goods. She appeals to the relation $x = I^S - I^D$ as we do, but I^D, though positive in her numerical runs, is used nowhere in the economy.

The revision above generates the numerical outputs reported in Appendix A.4 of Chichilnisky (1983). We present one run using our computer program in Table A1. These results are identical to those in her A.4, "Run 2". We present a computer generated graph in Figure A1. Note the fact that her GNP, the measure of aggregate output in the oil importing country, rises over a range of rising oil prices (her "sustitution effect") and then declines (her "income effect").

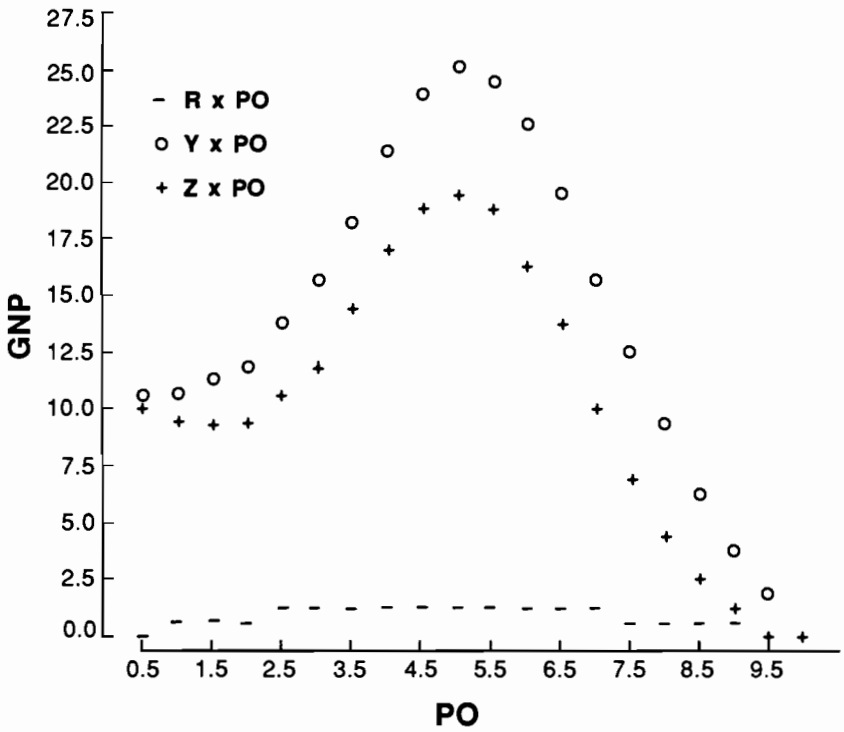


Figure A1 Oil Price Rises in the Chichilnisky Model

Table A1 Oil Price Rises in the Chichilnisky Model

A1 = 0.30 A2 = 0.20 B1 = 0.10 B2 = 0.20 C1 = 0.001 C2 = 0.60
ALPHA = 1.00 BETA = 2.00 PB = 1.00 GAMMA = 0.10

PO	W	R	IS	ID
0.5000	3.1667	0.2375	0.7917	0.1128
1.0000	3.0000	0.4500	1.5000	0.4050
1.5000	2.8333	0.6375	2.1250	0.8128
2.0000	2.6667	0.8000	2.6667	1.2799
2.5000	2.5000	0.9375	3.1249	1.7576
3.0000	2.3333	1.0500	3.5000	2.2050
3.5000	2.1667	1.1375	3.7917	2.5879
4.0000	2.0000	1.2000	4.0000	2.8800
4.5000	1.8333	1.2375	4.1250	3.0627
5.0000	1.6667	1.2500	4.1666	3.1249
5.5000	1.5000	1.2375	4.1251	3.0629
6.0000	1.3333	1.2000	4.0001	2.8801
6.5000	1.1667	1.1375	3.7917	2.5878
7.0000	1.0000	1.0500	3.5000	2.2050
7.5000	0.8333	0.9375	3.1251	1.7578
8.0000	0.6667	0.8000	2.6667	1.2800
8.5000	0.5000	0.6375	2.1250	0.8128
9.0000	0.3333	0.4500	1.5000	0.4050
9.5000	0.1667	0.2375	0.7917	0.1128
10.0000	0.0000	0.0000	0.0000	0.0000

PO	TRADE SURPLUS	PIX	Z	X
0.5000	0.0000	0.5806	10.1243	0.6789
1.0000	0.0000	1.2000	9.4438	1.0950
1.5000	0.0000	1.8417	9.1685	1.3122
2.0000	0.0000	2.4889	9.4084	1.3867
2.5000	0.0000	3.1250	10.2673	1.3673
3.0000	0.0000	3.7333	11.8010	1.2950
3.5000	0.0000	4.2973	13.9324	1.2038
4.0000	0.0000	4.8000	16.3429	1.1200
4.5000	0.0000	5.2250	18.4267	1.0622
5.0000	0.0000	5.5555	19.4430	1.0417
5.5000	0.0000	5.7746	18.9025	1.0621
6.0000	0.0000	5.8664	16.8638	1.1200
6.5000	0.0000	5.8139	13.8587	1.2039
7.0000	0.0000	5.6000	10.5349	1.2950
7.5000	0.0000	5.2084	7.3908	1.3673
8.0000	0.0000	4.6222	4.7111	1.3867
8.5000	0.0000	3.8251	2.6194	1.3122
9.0000	0.0000	2.8000	1.1467	1.0950
9.5000	0.0000	1.5306	0.2821	0.6789
10.0000	0.0000	0.0000	0.0000	0.0000

REFERENCES

Chichilnisky, G., 1983, "Oil Prices, Industrial Prices and Outputs: A General Equilibrium Macro Analysis", mimeo. (There are different versions of this paper circulating. We worked with one with detailed Appendices. One such version is "Prix du pétrole, prix industriels et production: une analyse macroéconomique d'équilibria général" in G. Gaudet and P. Lasserre (eds.), *Ressources Naturelles et Théorie Economique*, pp. 25-56, Les Presses de L'université Laval, Quebec, 1986).

Corden, W.M. and J.P. Neary, 1982, "Booming Sector and De-industrialization in a Small Open Economy", *Economic Journal* 92:285-99, June.

Corden, W.M., 1983, "The Economics of a Booming Sector", *International Social Science Journal* 35, 3:441-454.

Hartwick, J.M., 1984, "Primary Producing versus Industrial Regions in International and Interregional Trade", presented at the International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria, October 1984.

Advances in Spatial Theory and Dynamics
Å.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 7

Direct Equilibria of Economies and Their Perfect Homogeneity Limits

B. Dejon, B. Güldner and G. Wenzel

1. INTRODUCTION AND OVERVIEW

In the seventies stochastic choice theory became increasingly important in the realms of traffic theory and location analysis; see, e.g., McFadden (1973; 1978), Daganzo and Sheffi (1977). Dejon and Güldner (1985) extended the scope of applicability of stochastic choice theory to cover the choice of prices - as well as quantities - by economic agents in the context of a location and land use model. Güldner (1984) addressed the problem of economic general equilibrium modelling by stochastic choice methods in a nonspatial framework. Dejon (1986) combined ideas from Dejon and Güldner (1985) and Güldner (1984) to outline a stochastic choice model of a rudimentary spatial economy.

The single most important concept in extending stochastic choice theory to cover the choice of prices - as well as quantities and other entities - is the concept of relative demand at the various price levels of markets, introduced by Dejon and Güldner (1985). Relative demand at some price level is the quotient of purchases by supply at that price level. By adding penalty terms that depend on relative demand at the various price levels involved in an economic agent's activity (called 'alternative' later on in this paper) to the original neoclassical utility functions of economic agents, one prevents supply and demand from drifting too far apart at any of the markets. Thus (stochastic) general equilibrium of an economy may be defined without recourse to explicitly postulating equality of (expected) supply and demand at the various markets. This is the reason for subsequently referring to the 'direct' equilibrium of an economy.

In order to point out that there are connections between non-cooperative game theory and stochastic equilibrium theory we have chosen to speak of weak Nash instead of stochastic equilibria. It may be interesting to note that there are two different ways - an elementary one, let us say, and a nonelementary one - of viewing stochastic equilibrium as a generalization of Nash equilibrium. The elementary way shall be sketched in Section 2, the main task of which is to introduce the basic notions for later developments.

The nonelementary way is described by Theorem 8 and rests on the notion of perfect homogeneity equilibrium. Perfect homogeneity of a population of economic agents prevails when all agents behave perfectly alike. Within our modelling framework this is achieved by assuming that all agents behave on the basis of a utility function that is the same for all. Thus, while in the nonperfect homogeneity case the utility functions of economic agents are conceived of as realizations of some random utility function, in the perfect homogeneity case there is no randomness any more; there is a single deterministic utility function to all agents of any single population.

The notion of perfect homogeneity equilibrium applies not only in economic general equilibrium modelling, but also in technically simpler situations like route choice modelling in traffic equilibrium theory. This is briefly alluded to in Remark 2 of Section 4. In economic general equilibrium modelling, technical difficulties in passing to the perfect homogeneity limit derive from some discontinuity of relative demand as a function of supply and demand; see Remark 1 of Section 4.

In a route choice context, Theorem 8 describes passage from stochastic to deterministic traffic equilibrium. The latter may be characterized in the following way: The route choices of trip makers constitute a deterministic traffic equilibrium if, under the prevailing traffic pattern, no trip maker has any incentive to unilaterally change his route. (This is what is commonly called a 'user optimizing' equilibrium; see, e.g., Dafermos (1971)). In other words: No substitute (different route) available to an economic agent (trip maker) appears under the given state of the system (traffic pattern) more attractive than the presently adopted alternative. The various routes chosen by the trip makers of any single origin-destination pair thus belong to the same optimal indifference set of these trip makers - who form a population in the sense introduced in Section 2. According to Theorem 8 that idea carries over to more general situations.

In economic general equilibrium theory the optimal indifference sets adopted, under perfect homogeneity equilibrium conditions, by economic agents of the various populations possess additional and specific features relating to the eminent role of prices. Thus, e.g., Theorem 9 says that, given some perfect homogeneity equilibrium, at each nonempty market supply and demand concentrate at a single price level, or else, at two neighboring ones. Two more specific features are quantity optimization by economic agents for parametric prices, and certain forms of market clearing.

At the sequel of Theorem 9 in Section 4, quantity optimization behavior is obtained under a so-called negligible rationing condition. The ensuing question of passage from individual to aggregate demand is addressed in Remark 6, using an example of a simplified shopping model. The outcome hinges upon the ease of obtaining the total number of economic agents constituting the population of shoppers under study, characterized, e.g., by zone of residence and available income.

In so-called bilevel models of an economy the sizes of populations are endogenous variables. They are studied at what may be called the upper modelling level, while at the lower level economic agents' choices of prices and quantities in various markets are being emulated. The bilevel structure of an economy is briefly alluded to at the end of Section 2.

Market clearing is briefly discussed at the end of Section 4. For storable goods, market clearing is introduced as stationarity of stock: supply from current production equals transactions (which serve to offset, among others, depreciation of stocks). As to labor and services (nonstorable commodities), in direct or perfect homogeneity (p.h.) equilibrium purchases are typically less than current supplies. However, as equilibrium prevails, it still appears expedient to speak of market clearing, thus introducing more or less explicitly some 'natural' rate of unemployment of labor or service resources, respectively.

Section 3 is mainly technical and serves to prove so-called exponential decay theorems that are needed in order to prove Theorem 9 (while Theorem 8 could well have been stated and proven at the end of Section 2).

2. WEAK NASH AND DIRECT EQUILIBRIA

The headline topics of this section shall be reviewed succinctly. A more detailed presentation may be found in Dejon (1986).

N denotes a finite collection of behaviorally nearly homogeneous populations of economic units (households, firms, etc.). To each population $n \in N$, there is a finite choice

set (or set of alternatives), denoted by A^n. (In game theoretic terminology alternatives would be called strategies.) $P^n(b)$ designates the number of economic units from population n that are exercising alternative $b \in A^n$ and will be called the occupation number of b (a denomination taken from Markov chain theory). With the assumption that each economic unit is exercising exactly one of its alternatives, the identity

$$\sum_{b \in A^n} P^n(b) = P^n$$

holds, where P^n designates the total number of economic units in population n. By $\underline{P}$ we indicate $(P^n(b))_{n \in N, b \in A^n}$, the vector of all occupation numbers. $\underline{P}$ will take the role of state vector of the system.

The choices of economic units are guided by utility functions. Let $\tilde{u}_i^n(b|\underline{P})$ denote the utility which the economic unit $i \in n$ attributes to alternative $b \in A^n$, given the state $\underline{P}$ of the system. (In order to be able to account for external effects in economic applications, the utility values of alternatives are posited state dependent.) Let $\tilde{P}^n(b|\underline{P})$ denote the number of economic units $i \in n$ that consider alternative b to be optimal with respect to their respective individual utility functions $\tilde{u}_i^n(\cdot|\underline{P})$ (ties being excluded, for simplicity's sake). $\tilde{P}^n(b|\underline{P})$ will be called the *voted occupation number* of b, as opposed to the actual occupation number $P^n(b)$. A state $\underline{P}$ is called a *weak Nash equilibrium state* of the system if $P^n(b) = \tilde{P}^n(b|\underline{P})$ for all $n \in N$ and $b \in A^n$, that is to say, actual occupation numbers equal voted ones.

In the particular situation where each population n consists of a single economic unit, i.e. $P^n = 1$ for all n, a weak Nash equilibrium state $\underline{P}$ constitutes an ordinary Nash equilibrium in the following sense: Each player (i.e. economic unit) judges the strategy (i.e. alternative) actually adopted as the best one possible given the choices of all other players (as specified by the state vector $\underline{P}$). Note: if any of the players would not judge so, he would not give his vote to the strategy adopted, whence the voted occupation number of that strategy would be zero while the actual one is one.

For the sake of precision, let us observe that the preceding interpretation of Nash equilibrium does not fully agree with common usage of the term in noncooperative game theory. The more common version would read as follows: A state $\hat{\underline{P}}$ of the system is a Nash equilibrium state if any single player, after switching to a strategy different from the one presently chosen, would find that he has not improved on his previous choice (under the usual proviso that all other players stay with their given choices). This rather subtle distinction in the understanding of Nash equilibrium was discussed by Dafermos (1971) in the context of the route assignment problem of traffic equilibrium theory. See Remark 2 in Section 4 for more detail on formulating the route assignment problem as a Nash equilibrium problem.

As it would be a hopeless task to try to identify all individual utility functions $\tilde{u}_i^n$, one makes the following *stochastic utility assumption*: There exists, for each $n \in N$, some 'strict utility function' $u^n(\cdot|\underline{P})$ such that the

$$\varepsilon_i^n(\cdot|\underline{P}) := \tilde{u}_i^n(\cdot|\underline{P}) - u^n(\cdot|\underline{P}), \quad i \in n,$$

may be reasonably treated as realizations of a random vector $\mathcal{E}^n(\cdot|\underline{P})$, reasonable in the sense that

$$\tilde{P}^n(b|\underline{P}) \approx P^n \cdot d^n(b|\underline{u}) \quad (\text{for } b \in A^n,\ n \in N)$$

where $d^n(b|\underline{u})$ denotes the probability that, for each $a \in A^n$,

$$\mathcal{E}^n(b|\underline{P}) + u^n(b|\underline{P}) \geq \mathcal{E}^n(a|\underline{P}) + u^n(a|\underline{P}).$$

Under certain assumptions about the distributions of the random vectors $\mathcal{E}^n(\cdot|\underline{P})$, the probabilities $d^n(b|\underline{u})$ may be calculated analytically. Thus, e.g.,

$$d^n(b|\underline{u}) \propto \exp\,(\gamma^n\, u^n(b|\underline{P})),$$

if the components of $\mathcal{E}^n$ are independently and identically extreme value distributed, with spread parameter γ^n. A more flexible formula for d^n is obtained on the basis of the generalized extreme value distribution for the components of $\mathcal{E}^n$; see MacFadden (1978).

Direct equilibrium shall be introduced as a weak Nash equilibrium in the particular context of economic general equilibrium modelling. Various types of economic systems can be distinguished by their particular types of choice sets A^n and related strict utility functions $u^n(\cdot|\cdot)$. Another interesting feature of most economic systems is that one may distinguish *two (sometimes even more) levels of choice sets*, which we shall label lower level and upper level, respectively.

For the description of lower level choice sets, a few notions have to be introduced first: We consider finitely many commodity markets $g \in G$. For each market, there is a discrete price-quantity plane (P-Q-plane) $(\mathbf{P}^g \times \mathbf{Q}^g) \cup 0^g$, $\mathbf{P}^g$ denoting the (discrete) price axis of market g, $\mathbf{Q}^g$ its (discrete) quantity axis, and 0^g the null quantity. $\mathbf{P}^g$ and $\mathbf{Q}^g$ are finite subsets of the real axis $\mathbf{R}$, the elements of $\mathbf{Q}^g$ all being larger than zero. $\mathbf{P}^g \times \mathbf{Q}^g$ denotes the Cartesian product of $\mathbf{P}^g$ and $\mathbf{Q}^g$, i.e. the set of all price-quantity pairs (p^g, q^g), $p^g \in \mathbf{P}^g$ and $q^g \in \mathbf{Q}^g$. The null quantity 0^g is being handled separately as it will not require any pairing with some price p^g.

Lower level choice sets A^n are subsets of the Cartesian product $\underset{g \in G}{\times} [(\mathbf{P}^g \times \mathbf{Q}^g) \cup 0^g]$. A population n is said not to participate in market g if the g-component of each of the alternatives $b \in A^n$ is 0^g; otherwise the population is said to *participate in market* g. Let N^g denote the set of all such populations. N^g is partitioned into two nonempty and disjoint subsets N^{gS} and N^{gD}. The populations $n \in N^{gS}$ act as suppliers at market g, while populations $n \in N^{gD}$ act as demanders at that market.

For the specification of lower level strict utility functions the notion of relative demand will play a crucial role. (Absolute) demand and (absolute) supply will have to be introduced first. Notice: demand, in this context, is always meant to be satisfied demand, i.e. actual purchases. Demand of commodity g, in physical units, by population $n \in N^{gD}$, at price level p^g is

$$D^n(p^g) := \sum_{b \in A p^g} q^g(b)\, P^n(b), \tag{1}$$

where Ap^g denotes the set of all those alternatives $b \in A^n$ the g-th price component of which is p^g, while $q^g(b)$ denotes the g-th quantity component of b. Aggregate demand of commodity g, at price level p^g is

$$D(p^g) := \sum_{n \in N^{g_D}} D^n(p^g). \tag{2}$$

As to supply, we distinguish supply from stock, $SS^n(p^g)$, of commodity g at price level p^g, by population $n \in N^{g_S}$, and supply from current production, $SC^n(p^g)$. Interpreting q^g as a time rate of production, one obtains

$$SC^n(p^g) = \sum_{b \in Ap^g} q^g(b)\, P^n(b), \tag{3}$$

with the same meaning of Ap^g and $q^g(b)$ as further above in the context of equation (1). ($D^n(p^g)$ and $D(p^g)$, by the way, are also time rates.) In analogy to (2), one introduces aggregate supplies from stock and from current production, respectively:

$$SS(p^g) := \sum_{n \in N^{g_S}} SS^n(p^g), \tag{4}$$

$$SC(p^g) := \sum_{n \in N^{g_S}} SC^n(p^g). \tag{5}$$

Aggregate total supply is $S(p^g) := SS(p^g) + SC(p^g)$. Finally, then, *relative demand* for commodity g, at price level p^g, is

$$r^g(p^g) := \begin{cases} D(p^g)/S(p^g) & \text{for } S(p^g) \neq 0 \\ 0 & \text{for } S(p^g) = 0. \end{cases} \tag{6}$$

As $D(p^g)$ is satisfied demand at price level p^g, the inequality $D(p^g) \leq S(p^g)$ necessarily holds, implying $r^g(p^g) \leq 1$.

We are now in a position to introduce the type of lower level strict utility functions $u^n(\cdot|\underline{P})$ that are specific for the direct equilibrium approach:

$$u^n(b|\underline{P}) = u_0^n(b) - \sum_{g \in G^b} z^{ng}\,(r^g(p^g(b))). \tag{7}$$

Here $u_0^n(\cdot)$ may, in principal, be any utility function which is common in neoclassical theories when external effects are not being accounted for; $u_0^n(\cdot)$ will be called the *original utility function* of population n. (Note: its argument is b, which comprises quantity as well as price components.) G^b is the set of all those markets where alternative b is active, and $z^{ng}(\cdot)$ is a *penalty function*, for population n at market g. If n is a supplier population at that market, $z^{ng}(r^g(p^g))$ penalizes price level p^g whenever relative demand $r^g(p^g)$ is low, and conversely, if population n is on the demand side of market g, $z^{ng}(r^g(p^g))$ penalizes such price levels where $r^g(p^g)$ is high, i.e. near to 1. On the part of demanders,

these penalties are interpretable as *disutilities of search*; on the part of suppliers, the appropriate interpretation appears to be *disutility of uncertainty of demand.*

While original utility functions are normally such that suppliers at any market g prefer higher price levels and demanders lower ones, the penalty terms z^{ng} in (7) will induce suppliers and demanders to avoid price levels where relative demand is rather small or close to 1, respectively.

Upper level choice sets are loosely circumscribed by stating that typically some or all of the alternatives of an upper level choice set are constituted by lower level populations. To elucidate this point, consider, e.g., the choice of location by firms. On the lower level, let there be populations of firms, P^1, P^2,..., P^k, say, in geographic areas Z^1,..., Z^k, respectively, these firms all being alike in all respects, except for their location. The (lower level) decisions of these firms consist of choosing sales strategies and production plans or, more precisely, prices and quantities. By contrast, locational decisions of these firms are modelled as choices in an upper level choice set A^0, the elements of which are the geographic areas, Z^1,..., Z^k, or, formally equivalent, the populations P^1,..., P^k. The population P^0 of economic units whose choice set is A^0 consists of the union of all populations P^1,..., P^k : $P^0 = \cup_\kappa P^\kappa$.

Upper level strict utility functions typically make use of lower level information about prices and quantities. We shall not dwell on this matter as the main focus of the remaining parts of this paper will be on lower level issues.

To sum up: The constitutive elements of a weak Nash equilibrium system are the following:

A finite set, N, of (finite) populations, each endowed with

β1) a (finite) set A^n of alternatives,

β2) a strict utility function $u^n(b|\underline{P})$,

β3) a distribution function $d^n(b|\underline{u},\gamma^n)$.

In economic general equilibrium modelling, some of the sets of alternatives typically consist of price-quantity pairs indexed by commodities. Weak Nash equilibria are then more specifically called *direct equilibria.* The epithet 'direct' was briefly commented upon in Section 1. In Dejon (1986), a brief outline may be found of a weak Nash equilibrium system representing some spatial economy.

3. QUALITATIVE ANALYSIS OF DIRECT EQUILIBRIA

In the second paragraph of the previous section we stated that we were going to deal with "nearly homogeneous" populations of economic units. This meant that their individual utility functions where not all the same, but could be treated as a random sample of a stochastic vector $u^n(\cdot|\underline{P}) + \mathcal{E}^n(\cdot|\underline{P})$ (where $u^n(\cdot|\underline{P})$ was non-random). The variances of the components of the random vector $\mathcal{E}^n$ may be taken as measures of the deviation of population n from perfect homogeneity. In the sequel, we shall express by a single parameter, γ^n, tending to infinity that the deviation of population n from perfect homogeneity tends to zero. During that process the distributions $d^n(\cdot|\underline{u},\gamma^n)$ grow

increasingly peaked; see the later concentrability assumption for a precise statement. The main focus of this section shall be on the behavior of direct equilibria as γ^n becomes very large.

The results to be achieved depend crucially on the penalties $z^{ng}(r^g(p^g(b)))$. We shall repeatedly deal with prices p^g at some market g where for some population n the penalty $z^{ng}(r^g(p^g(b)))$ is larger than zero. This will then be expressed by saying that population n is *rationed* at price level p^g of market g.

A first assumption to be made is one of *monotonicity of the penalty functions* z^{ng} as functions of relative demand: Depending on whether population n is on the supply or on the demand side of market g, the penalty function $z^{ng}(r^g)$ increases or decreases, respectively, as relative demand decreases.

A similarly natural postulate is that of *monotonicity of the original utility functions* with respect to prices (notwithstanding Veblen's 'conspicuous consumption'). To express this formally, we need some notation: For any alternative $b \in A^n$, $n \in N$, we designate by $b\neg p^g$, and by $b\neg p^g, q^g$ the alternatives one obtains after replacing the g-th price coordinate and the g-th price-quantity coordinate of b by p^g and p^g, q^g, respectively. Furthermore, $p^g <_n p'^g$ means:

$$p'^g < p^g \quad \text{if} \quad n \in N^{gD}, \text{ and}$$

$$p'^g > p^g \quad \text{if} \quad n \in N^{gS}.$$

Monotonicity of the u_0^n now reads:

$$p^g <_n p'^g \Rightarrow u_0^n(b\neg p^g) \leq u_0^n(b\neg p'^g)$$

for all $b \in A^n$ with $q^g(b) \neq 0^g$.

Last, but not least we posit *monotonicity of the distribution functions* $d^n(b|\underline{u}, \gamma^n)$ in the following sense:

$$u^n(b') \geq u^n(b) \Rightarrow d^n(b'|\underline{u}, \gamma^n) \geq d^n(b|\underline{u}, \gamma^n) \quad (b, b' \in A^n).$$

We may now prove a further monotonicity property.

Theorem 1: In any direct equilibrium state $\underline{P}$ of an economy, at any market g, relative demand is a nonincreasing function of price: $p^g > p'^g \Rightarrow r^g(p^g) \leq r^g(p'^g)$.

Proof (by contradiction): Assume there are two price levels $p^g > p'^g$ with $r^g(p^g) > r^g(p'^g)$. Then for any population of suppliers $n \in N^{gS}$ and any alternative $b \in A^n$ with $q^g(b) > 0$:

$$u^n(b\neg p^g) \geq u^n(b\neg p'^g)$$

and, thence, $d^n(b\neg p^g|\underline{u}, \gamma^n) \geq d^n(b\neg p'^g|\underline{u}, \gamma^n)$ with $\underline{u} := (u^n(b|\underline{P}))_{n \in N, b \in A^n}$. Therefore, also $q^g \, P^n \, d^n(b\neg p^g, q^g|\underline{u}, \gamma^n) \geq q^g \, P^n \, d^n(b\neg p'^g, q^g|\underline{u}, \gamma^n)$.

By summing on both sides of this inequality for fixed b over all $q^g \in Q^g$, and subsequently over all $b \in A^n$ in as much as they differ at some market g' other than g, one obtains

$$SC^n(p^g) \geq SC^n(p'^g). \tag{8}$$

Summing in addition over all $n \in N^{gS}$ yields $SC(p^g) \geq SC(p'^g)$.

We now introduce the following *proportional total supply assumption*: At any price level p^g, total supply, $S(p^g)$, at market g is proportional to supply from current production: $S(p^g) = k^g\, SC(p^g)$, k^g some constant larger than or equal to unity. The preceding inequality then immediately implies $S(p^g) \geq S(p'^g)$. In a similar vein, one deduces for total demands, i.e. purchases: $D(p^g) \leq D(p'^g)$. The last two inequalities imply $r^g(p^g) \leq r^g(p'^g)$, contradicting the inequality we started out from.

Remark: The proportionality constant k^g in the above proportional total supply assumption would more plausibly be posited to depend on population n, too, i.e. to be k^{ng} instead of k^g. This, however, would render the proofs of Theorems 4 and 6 more involved.

As our next assumption we introduce *complementarity of penalty functions* z^{ng}: To any market g there is a number α^g, $0 < \alpha^g < 1$,such that no supplier is rationed at any price level p^g with $r^g(p^g) \geq \alpha^g$, and no demander is rationed at any price level p^g with $r^g(p^g) \leq \alpha^g$. (While this assumption may well appear to be too restrictive, it simplifies subsequent reasoning. Relaxation of the assumption will have to be considered elsewhere.)

With Theorem 1 in mind, we set p^{gS} equal to the highest price level at market g at which no supplier is rationed (provided such a price level exists), and similarly p^{gD} equal to the lowest price level at market g at which no demander is rationed (provided such a price level exists).

Notice: p^{gS} cannot be "much smaller" than p^{gD} in the sense that there does not exist $p^g \in P^g$ such that $p^{gS} < p^g < p^{gD}$. For at price level p^g at least one $n \in N^{gS}$ and at least one $n \in N^{gD}$ would be rationed in contradiction to the above complementarity assumption.

We shall soon prove that under certain conditions p^{gS} cannot be larger than p^{gD}. To this end, we shall need the following assumption of *concentrability* of the functions d^n: For any $b \in A^n$ and $\Delta > 0$ let $M^n_\Delta(b) := \{a \in A^n \mid u^n(a) \leq u^n(b) - \Delta\}$ and $d^n(M^n_\Delta(b) \mid \underline{u}, \gamma^n) := \sum_{a \in M^n_\Delta(b)} d^n(a \mid \underline{u}, \gamma^n)$. The concentrability assumption then says: For any $\rho > 0$, $\Delta > 0$ there exists $\gamma(\rho, \Delta)$ such that $d^n(M^n_\Delta(b) \mid \underline{u}, \gamma^n) \leq \rho\, d^n(b \mid \underline{u}, \gamma^n)$ for all $\gamma^n \geq \gamma(\rho, \Delta)$. This means that for sufficiently large γ^n the voted occupation numbers decay exponentially with decreasing strict utility values of alternatives. In the case of generalized multinomial logit models, e.g., this kind of concentrability prevails indeed.

In order to prove $p^{gS} \leq p^{gD}$ (for sufficiently large γ^n) we need, in addition to concentrability, the following assumption of *strong monotonicity of the original utility functions* u^n_0: There exists $\underset{\sim}{\Delta} > 0$ such that

$$\underset{\sim}{\Delta} + \max_{qg} u_0^n(b\neg pg,qg) \leq \max_{qg} u_0^n(b\neg p'g,qg) \tag{9}$$

for all $n\in N$, $b\in A^n$, provided $pg <_n p'g$. In many instances, inequality (9) is checked by showing that

$$\underset{\sim}{\Delta} + u_0^n(b\neg pg,qg) \leq u_0^n(b\neg p'g,qg) \tag{10}$$

holds for all qg, provided $pg <_n p'g$. This simple procedure does not work out, however, if one deals, e.g., with an (original) utility function u_0^h for consumers that depends only on quantities consumed, but not on prices (and which takes the value $-\infty$ outside the budget set). For in that situation, $u_0^n(b\neg pg,qg) = u_0^n(b\neg p'g,qg)$ whenever both utility values are finite (i.e., $b\neg pg,qg$ and $b\neg p'g,qg$ are both within the budget set). Inequality (9), however, will hold if there exists some quantity $q'g\in Qg$ that large that $b\neg pg,q'g$ lies outside the budget set while $b\neg p'g,q'g$ still lies inside, and if the utility decrease from $b\neg p'g,q'g$ to any $b\neg p'g,q"g$, $q"g < q'g$, amounts to at least $\underset{\sim}{\Delta}$. If consumers attribute some utility to savings, the original utility function u_0^h depends on quantities consumed as well as on prices paid and, normally then, inequality (10) should hold.

Theorem 2: In any direct equilibrium state $\underline{P}$ of an economy, at any market g, in the nonrationed price range of any single population n, the occupation numbers P^n_{pg} at price levels pg decay exponentially with worsening price levels pg, provided γ^n is sufficiently large. More precisely: $pg <_n p'g \leq_n pg^S$ or pg^D, respectively, implies $P^n_{pg} \leq \rho\, P^n_{p'g}$ for $\gamma^n \geq \gamma(\rho, \underset{\sim}{\Delta})$.

Proof: In equilibrium, actual occupation numbers $P^n(b)$ equal voted occupation numbers $P^n\, d^n(b|\underline{u}, \gamma^n)$. For any $b\in A^n$ let $\hat{b}$ denote that alternative in the set $\{b\neg p'g,qg \mid qg\in Qg\}$ which is optimal with respect to original utility u_0^n. Because of strong monotonicity $\{b\neg pg,qg \mid qg\in Qg\} \subseteq M^n_{\underset{\sim}{\Delta}}(\hat{b})$. Therefore,

$$\begin{aligned}\sum_{qg} P^n(b\neg pg,qg) \leq P^n\,(M^n_{\underset{\sim}{\Delta}}(\hat{b})) &= P^n\, d^n(M^n_{\underset{\sim}{\Delta}}(\hat{b})|\underline{u},\gamma^n) \leq \rho\, P^n\, d^n(\hat{b}|\underline{u},\gamma^n) \\ &= \rho\, P^n(\hat{b}) \leq \rho \sum_{qg} P^n(b\neg p'g,qg)\end{aligned}$$

for $\gamma^n \geq \gamma(\rho, \underset{\sim}{\Delta})$ (concentrability!).

Summing both sides of the inequality over all $b\in A^n$ in as much as they differ at some market g' other than g one finally obtains $P^n_{pg} \leq \rho\, P^n_{p'g}$ for any $\gamma^n \geq \gamma(\rho, \underset{\sim}{\Delta})$.

Theorem 2 will be called the *single population exponential decay theorem*. By summation over all populations $n\in Ng^S$, and $n\in Ng^D$ in turn, one obtains what will be called the *all suppliers* or *all demanders exponential decay theorem*, respectively:

Theorem 3: In any direct equilibrium state $\underline{P}$ of an economy, at any market g, the summed occupation numbers $P^S_{pg} := \sum_{n \in N^{gS}} P^n_{pg}$ decay exponentially with worsening price levels $p^g \leq p^{gS}$, provided all γ^n, $n \in N^{gS}$, are sufficiently large. An analogous statement holds for the summed occupation numbers $P^D_{pg} := \sum_{n \in N^{gD}} P^n_{pg}$ in the range $p^g \geq p^{gD}$.

As demand $D(p^g)$ at any price level p^g is forcedly not larger than (total) supply $S(p^g)$ the all suppliers exponential decay theorem induces an all demanders exponential decay theorem in the price range $p^g \leq p^{gS}$. For its formulation some more notation is needed.

Let

$$q^g_{max} := \max q^g \in Q^g,$$

$$q^g_{min} := \min q^g \in Q^g > 0 \text{ (!)},$$

and $\tilde{\alpha}^g$ (>0) equal the largest value of relative demand below which each population of suppliers is rationed.

Theorem 4: In any direct equilibrium state $\underline{P}$ of an economy, at any market g, and for any two price levels $p^g < p'^g \leq p^{gS}$ $P^D_{pg} \leq \hat{\rho}\, P^D_{p'g}$ with

$$\hat{\rho} := \frac{\rho}{\tilde{\alpha}^g} \left(\frac{q^g_{max}}{q^g_{min}} \right)^2, \text{ provided } \gamma^n \geq \gamma(\rho, \underset{\sim}{\Delta}) \text{ for all } n \in N^{gS} \cup N^{gD}.$$

Notice: As ρ can be chosen arbitrarily small, so can $\hat{\rho}$.

Proof of Theorem 4:

$$\frac{P^D_{pg}}{P^D_{p'g}} \leq \frac{q^g_{max}\, D(p^g)}{q^g_{min}\, D(p'^g)} = \frac{q^g_{max}\, r^g(p^g)\, S(p^g)}{q^g_{min}\, r^g(p'^g)\, S(p'^g)}$$

$$= \frac{q^g_{max}\, r^g(p^g)\, SC(p^g)}{q^g_{min}\, r^g(p'^g)\, SC(p'^g)} \qquad \text{(cf. proportional total supply assumption)}$$

$$\leq \left(\frac{q^g_{max}}{q^g_{min}} \right)^2 \frac{r^g(p^g)}{r^g(p'^g)} \frac{P^S_{pg}}{P^S_{p'g}} \leq \left(\frac{q^g_{max}}{q^g_{min}} \right)^2 \frac{1}{\tilde{\alpha}^g}\, \rho.$$

Theorem 4 will be called the *secondary all demanders exponential decay theorem*; secondary, because it is derived from a "primary" exponential decay theorem, viz., the one for suppliers.

We have already seen that pg^S cannot be "much smaller" than pg^D. In addition, we can now state

Theorem 5: In any direct equilibrium state $\underline{P}$ of an economy, at any market g, it cannot happen that $pg^S > pg^D$, whenever ρ is chosen smaller than $\tilde{\alpha}^g(q^g_{min}/q^g_{max})^2$ and $\gamma^n \geq \gamma(\rho, \underset{\sim}{\Delta})$ for all $n \in Ng^S \cup Ng^D$.

Proof (by contradiction): Notice first that the assumption made about ρ means that not only $\rho < 1$, but also $\hat{\rho} < 1$. Assume now $pg^S > pg^D$. Necessarily then $r^g(pg^D) > 0$ and, therefore, also $D(pg^D) > 0$. For $D(pg^D) > 0$ Theorem 3 says that P^D_{pg} decreases strictly as pg ($\geq pg^D$) increases. On the other hand, Theorem 4 says that P^D_{pg} does not decrease as pg increases anywhere in the range $pg^D \leq pg \leq pg^S$. Thence, a contradiction!

Theorem 5, together with the observation that pg^S cannot be "much smaller" than pg^D, says that for sufficiently large values of the γ^n either $pg^S = pg^D$ or else pg^S is just one discrete price step below pg^D (provided both prices exist, i.e., provided no side of the market is rationed at all prices $pg \in Pg$).

In order to deduce a secondary all *suppliers* exponential decay theorem we have to make three more assumptions. The first is the following *bounded monotonicity assumption for suppliers' original utility*: There exists $\tilde{\Delta} > 0$ such that for any $g \in G$, $n \in Ng^S$

$$\tilde{\Delta} + u^n_0(b \neg pg, qg) \geq u^n_0(b \neg p'g, qg) \tag{11}$$

for all $b \in A^n$, $qg \in Qg$, whenever $p'g$ is just one discrete price step above pg.

Another assumption to be made is a *strong monotonicity assumption for the penalty functions* z^{ng} *of suppliers* $n \in Ng^S$ at market g: There exists $0 < \eta < 1$ such that any decrease in relative demand by at least $100\,\eta$ % (in the range where z^{ng} is larger than zero) induces an increase of z^{ng} by at least $\tilde{\Delta} + \underset{\sim}{\Delta}$; this is assumed to hold at all markets g for all $n \in Ng^S$.

The assumption implies that the maximal increase $\tilde{\Delta}$ of original utility on the part of any supplier when passing from one price level to the next higher one will be offset by at least $\underset{\sim}{\Delta}$ if under that price step relative demand decreases by at least $100\,\eta$ % (in the range where z^{ng} is larger than zero).

The last assumption to be made is a *no complementarity gap* assumption, and reads as follows: For any relative demand smaller than αg all suppliers at market g are rationed. This is assumed to hold for all markets. The assumption says that $\alpha g = \tilde{\alpha} g$ (see Theorem 4).

We are now in a position to prove the above-mentioned *secondary all suppliers exponential decay theorem*:

Theorem 6; In any direct equilibrium state $\underline{P}$ of an economy, at any market g, under the ten assumptions made previously (listed at the end of this section) the following holds: For any $p'^g > p^g > p^{gS}$

$$\frac{P^S_{p'g}}{P^S_{pg}} \leq \frac{\rho}{1-\eta}\left(\frac{q^g_{max}}{q^g_{min}}\right)^2 \qquad \text{provided } \gamma^n \geq \gamma(\rho, \underset{\sim}{\Delta}) \text{ for all } n \in N^{gS} \cup N^{gD}.$$

Proof: Case 1: $r^g(p'^g) \leq (1-\eta)\, r^g(p^g)$. As then $u^n(b_{\neg}p'^g) \leq u^n(b_{\neg}p^g) - \underset{\sim}{\Delta}$ (because of the strong monotonicity assumption for the penalty functions of suppliers in conjunction with the no complementarity gap assumption), concentrability provides for the inequality $P^n(b_{\neg}p'^g) \leq \rho\, P^n(b_{\neg}p^g)$ (whenever $\gamma^n \geq \gamma(\rho, \underset{\sim}{\Delta})$. Summing both sides of the inequality over all alternatives $b \in A^n$ (in as much as they differ not only by their price levels at market g) and also summing over all $n \in N^{gS}$ one obtains

$$P^S_{p'g} \leq \rho\, P^S_{pg} \qquad \text{(provided } \gamma^n \geq \gamma(\rho, \underset{\sim}{\Delta}) \text{ for all } n \in N^{gS}).$$

Case 2: $r^g(p'^g) > (1-\eta)\ r^g(p^g)$. This means

$$\frac{D(p'^g)}{S(p'^g)} > (1-\eta)\,\frac{D(p^g)}{S(p^g)}, \quad \text{or} \quad \frac{S(p'^g)}{S(p^g)} < \frac{1}{1-\eta}\,\frac{D(p'^g)}{D(p^g)}.$$

By the proportional total supply assumption then

$$\frac{SC(p'^g)}{SC(p^g)} < \frac{1}{1-\eta}\,\frac{D(p'^g)}{D(p^g)}.$$

Thence,

$$\frac{q^g_{min} P^S_{p'g}}{q^g_{max} P^S_{pg}} < \frac{1}{1-\eta}\,\frac{q^g_{max} P^D_{p'g}}{q^g_{min} P^D_{pg}} \leq \frac{\rho}{1-\eta}\,\frac{q^g_{max}}{q^g_{min}}, \quad \text{or}$$

$$\frac{P^S_{p'g}}{P^S_{pg}} < \frac{\rho}{1-\eta}\left(\frac{q^g_{max}}{q^g_{min}}\right)^2 \quad (> \rho).$$

Notice: The right hand side can be rendered arbitrarily small by choosing ρ sufficiently small.

With $P^{S \text{ resp.} D}_{\{pg > pg^D\}} := \sum_{pg > pg^D} P^{S \text{ resp.} D}_{pg}$ and an analogous definition of $P^{S \text{ resp. } D}_{\{pg < pg^S\}}$ one deduces from the primary and secondary exponential decay theorems (by use of the summation formula for the geometric series)

$$P^{S \text{ resp. } D}_{\{pg>pg^D\}} \leq \frac{\rho'}{1-\rho'} P^{S \text{ resp.} D}_{pg^D} \quad \text{and} \tag{12}$$

$$P^{S \text{ resp. } D}_{\{pg<pg^S\}} \leq \frac{\rho'}{1-\rho'} P^{S \text{ resp.} D}_{pg^S}, \tag{13}$$

where ρ' designates the largest of the two decay factors

$$\frac{\rho}{1-\eta}\left(\frac{q^g_{max}}{q^g_{min}}\right)^2 \text{ and } \frac{\rho}{\bar{\alpha}^g}\left(\frac{q^g_{max}}{q^g_{min}}\right)^2, \text{ and where, in addition, } \rho \text{ is so small that}$$

$\rho' < 1$. From (12) and (13) one obtains

Theorem 7: When all four exponential decay theorems hold and the preceding ρ' is less than 1, then the number of suppliers or demanders that find themselves on price levels pg^S and pg^D, whether these coincide or not, is more than $P^{g^S}_{act}(1-\rho')/(1+\rho')$ or $P^{g^D}_{act}(1-\rho')/(1+\rho')$, respectively. Here $P^{g^S}_{act}$ designates the total number of those suppliers at market g that are *active in that market*, i.e. who have not chosen $q^g := 0^g$. The analogous interpretation applies to $P^{g^D}_{act}$.

To conclude this section, it appears expedient to provide a synopsis of the ten assumptions made so far:

(A1) Monotonicity of the penalty functions z^{ng}:
$r^g <_n r'^g \Rightarrow z^{ng}(r^g) \geq z^{ng}(r'^g)$.

(A2) Monotonicity of original utility functions u^n_0 with respect to prices:
$p^g <_n p'^g \Rightarrow u^n_0(b \neg p^g) \leq u^n_0(b \neg p'^g)$.

(A3) Monotonicity of the distribution functions $d^n(b|\underline{u},\gamma^n)$ with respect to utilities:
$u^n(b) \leq u^n(b') \Rightarrow d^n(b|\underline{u},\gamma^n) \leq d^n(b'|\underline{u},\gamma^n)$.

(A4) Proportional total supply assumption:
$S(p^g) = k^g\, SC(p^g)$.

(A5) Complementarity of the penalty functions z^{ng}: For $r^g \leq \alpha^g$ no demander is rationed; for $r^g \geq \alpha^g$ no supplier is rationed.

(A6) Concentrability of the distribution functions $d^n(b|\underline{u},\gamma^n)$:
$\gamma^n \geq \gamma(\rho, \Delta) \Rightarrow d^n(M^n_\Delta(b)|\underline{u},\gamma^n) \leq \rho\, d^n(b|\underline{u},\gamma^n)$.

(A7) Strong monotonicity of the original utility functions u_0^n: For some $\underset{\sim}{\Delta} > 0$ and $pg <_n p'g$
$\underset{\sim}{\Delta} + \max_{qg} u_0^n(b\neg pg,qg) \leq \max_{qg} u_0^n(b\neg p'g,qg)$.

(A8) Bounded monotonicity of suppliers' original utility:
For some $\tilde{\Delta} > 0$ and $pg <_n p'g$ by just one discrete step
$\tilde{\Delta} + u_0^n(b\neg pg,qg) \geq u_0^n(b\neg p'g,qg)$.

(A9) Strong monotonicity of suppliers' penalty functions z^{ng}:
$z^{ng}((1-\eta)rg) - \tilde{\Delta} - \underset{\sim}{\Delta} \geq z^{ng}(rg)$ (> 0).

(A10) No complementarity gap: For relative demand $r^g < \alpha^g$ all suppliers are rationed.

4. PERFECT HOMOGENEITY EQUILIBRIA AND THEIR WALRASIAN PROPERTIES

We start out by introducing the following notion: Occupation vector $\underline{P}$ belongs to the (ρ,Δ)-*class* of some given vector $\underline{u}$ of utilities means

$$P^n(M_\Delta^n(b)) \leq \rho\, P^n(b) \text{ for all } n\in N,\ b\in A^n,$$

where $M_\Delta^n(b)$ is defined as above in connection with the concentrability assumption. That assumption, by the way, says that for $\gamma^n \geq \gamma(\rho,\Delta)$ $P^n d^n(\cdot|\underline{u},\gamma^n)$ belongs to the (ρ,Δ)-class of $\underline{u}$.

A state $\hat{\underline{P}}$ is defined to be a *perfect homogeneity equilibrium* (p.h. equilibrium) of the weak Nash equilibrium system $(N,(A^n),(u^n(\cdot|\cdot)),(d^n(\cdot|\cdot)))$ if for each population n there is a function v^n on A^n such that for any $\varepsilon > 0, \rho > 0, \Delta > 0$ there exists a weak Nash equilibrium state $\tilde{\underline{P}}$ with the following properties:

(i) $|\hat{P}^n(b) - \tilde{P}^n(b)| < \varepsilon$

(ii) $|v^n(b) - u^n(b|\tilde{\underline{P}})| < \varepsilon$ for all $n\in N, b\in A^n$,

(iii) $\tilde{\underline{P}}$ belongs to the (ρ,Δ)-class of $(u^n(\cdot|\tilde{\underline{P}}))_{n\in N}$.

In general terms this means that p.h. equilibria can be arbitrarily well approximated by weak Nash equilibrium states $\tilde{\underline{P}}$ that belong to the (ρ,Δ)-class of $(u^n(\cdot|\tilde{\underline{P}}))_{n\in N}$.

Remark 1: If at the 'point' $\hat{\underline{P}}$ the strict utilities $u^n(b|\underline{P})$, $n\in N, b\in A^n$, depend continuously on $\underline{P}$, then necessarily $v^n(b) = u^n(b|\hat{\underline{P}})$ for all $n\in N, b\in A^n$. Unfortunately, in economic general equilibrium modelling the utilities $u^n(b|\underline{P})$ do not depend continuously on $\underline{P}$ at 'points' $\hat{\underline{P}}$ that at some market $g\in G^b$ induce vanishing supply $S(p^g)$ at the price level p^g of alternative b. This pertains to the fact that relative demand as defined under (6) is, at the point $\widehat{D(p^g)} = \widehat{S(p^g)} = 0$, a discontinuous function of the two variables $D(p^g)$, $S(p^g)$. Note: The previous theorems, in particular Theorem 1, are not

hampered by this discontinuity of relative demand. Terminology: The values $v^n(b)$ in the above definition of a p.h. equilibrium $\hat{\underline{P}}$ are called *p.h. utilities pertaining to* $\hat{\underline{P}}$.

> **Theorem 8:** Let $\hat{\underline{P}}$ be a p.h. equilibrium of the weak Nash equilibrium system $(N,(A^n), (u^n(\cdot|\cdot)),(d^n(\cdot|\cdot)))$ with pertaining p.h. utilities $v^n(b)$. Then $\hat{\underline{P}}$ is a Nash equilibrium state in the following sense: $\hat{P}^n(b) > 0$ for some $n \in N, b \in A^n$ implies that b is optimal with respect to v^n, i.e. $v^n(b) \geq v^n(a)$ for all $a \in A^n$.

Proof (by contradiction): Assume there is some $a \in A^n$ which $v^n(a) - v^n(b) = 3\Delta > 0$. As $\hat{\underline{P}}$ is a p.h. equilibrium there exists for arbitrary $\rho > 0$ and $\varepsilon \leq \Delta$ a weak Nash equilibrium state $\tilde{\underline{P}}$ satisfying (i), (ii) and (iii) above. As (iii) implies $\tilde{P}^n(b) \leq \rho\ \tilde{P}^n(a)$, as furthermore $|\tilde{P}^n(a) - \hat{P}^n(a)| < \varepsilon$, $|\tilde{P}^n(b) - \hat{P}^n(b)| < \varepsilon$, and ρ as well as ε may be chosen arbitrarily small, one concludes that $\hat{P}^n(b)$ is arbitrarily small, i.e. equal to zero. This however, contradicts the assumption made in the Theorem.

Remark 2: The route assignment (or route choice) problem of traffic equilibrium theory may be conceived of as a p.h. equilibrium problem for a weak Nash equilibrium system of the following type: To each origin-destination pair, w, corresponds exactly one population, also designated by w, the economic units of which are formed by the trips (or the trip makers) to be assigned to routes connecting the origin of w to its destination. It is these routes that constitute the set A^w of alternatives of population w. A state vector $\underline{P}$ then is what is commonly called a traffic pattern. As strict utility functions $u^w(b|\underline{P})$ one adopts the negatives of some generalized trip cost functions $c^w(b|\underline{P})$. In traffic equilibrium theory, weak Nash equilibria are commonly called stochastic user equilibria, see e.g. Daganzo and Sheffi (1977). As a rule, generalized trip cost $c^w(b|\underline{P})$ depends continuously on the traffic pattern $\underline{P}$. Therefore, if $\hat{\underline{P}}$ designates some p.h. equilibrium of our route choice weak Nash equilibrium system, the pertaining p.h. utilities are $v^w(b) := -c^w(b|\hat{\underline{P}})$; note Remark 1. Thus, contrary to what happens in economic general equilibrium modelling, as for any possible p.h. equilibrium $\hat{\underline{P}}$ one can specify a more or less simple algebraic expression for $v^n(b)$, p.h. equilibrium can be calculated as the solution of the following 'user optimization problem' (as it is commonly called in traffic equilibrium theory; see, e.g., Dafermos (1971)): Determine that traffic pattern $\hat{\underline{P}}$, for which actually used routes are all optimal in the sense of exhibiting minimal (load dependent) generalized travel cost: $\hat{P}(b) > 0 \Rightarrow c^w(b|\hat{\underline{P}}) \leq c^w(a|\hat{\underline{P}})$ for all $a \in A^w$. This Nash equilibrium problem for a game with the individual trip makers as players might as well be recast as a game wherein the populations w are the players. As the payoff function of player w one would then have to adopt the sum of integrals over certain link flows as is common in network flow theory; see, e.g., Devarayan (1981) or Iri (1969). If, by the way, one adopts as the payoff function a sum of products of certain link flows by the respective (load dependent) individual link flow costs, one obtains a game whose Nash equilibria are, in traffic

equilibrium theory, commonly called 'system optimizing' flow patterns; see, e.g., Haurie and Marcotte (1985).

Remark 3: In the context of the route choice problem, Theorem 8 was proven by Müller (1980) for the particular case of distribution functions $d^n(\cdot|\underline{u},\gamma^n)$ as provided by the multinomial logit formula. Evans (1973) obtained an analogous result for the gravity model proving convergence towards a linear programming model of the trip distribution problem as dispersion in the gravity model tends to zero.

Remark 4: In economic general equilibrium modelling, the typical situation where Theorem 8 comes to bear occurs when economic agents possess substitution possibilities. In a spatial economics context a frequent case of substitutability is the one where physically alike commodities are available at different geographical locations. More specifically, consider some population of private households, $y\hat{r}$, residing in zone $\hat{r}$, of some metropolitan area. Let the consumption alternatives of that population consist of buying, in various geographical zones r, certain amounts of the local composite good gr, at a certain number of prices. For the sake of simplicity, let us assume that any single household shops only in a single zone per trip - which means that the quantity components q^{gr} of any consumption alternative b are all zero, except for one. (This type of model is outlined in greater detail in Dejon (1986)). The original utility of such an alternative could be of the form

$$u_0^{y\hat{r}}(b) := u_{0r}^{y\hat{r}}(p^{gr} + t^r, q^{gr})$$

where r designates that zone in which, under alternative b, shopping actually occurs, and t^r is the unit transport cost as incurred by residents of zone $\hat{r}$ when shopping in zone r. In p.h. equilibrium let there be, e.g., five zones $r_1,...,r_5$ where shopping by households from population $y\hat{r}$ actually occurs. This means that the set of those alternatives $b \in A^{y\hat{r}}$ that are actually occupied may be partitioned into five non empty subsets $B_{r1},...,B_{r5}$ according to their non-vanishing quantity components q^{gr}. Any two alternatives from any single subset B_{ri} differ by their price or quantity components at market gr_i, but are both zero at all other markets. As p.h. equilibrium prevails, the p.h. utilities $v^{y\hat{r}}(b)$ of the alternatives $b \in B_{r1} \cup ... \cup B_{r5}$ are all the same (and optimal) according to Theorem 8: $\hat{v} = v(b)$ for all $b \in B_{r1} \cup ... \cup B_{r5}$. By use of the continuity argument outlined in Remark 1 one obtains - for every $b \in B_{r1} \cup ... \cup B_{r5}$

$$v(b) = u_0^{y\hat{r}}(b) - z^{y\hat{r},gr}(r^{gr}(p^{gr}(b))),$$

if b means shopping in zone r. In Theorem 9 we shall see that the preceding disutility term $z^{y\hat{r},gr}$ is practically negligible for sufficiently fine mesh spacings of the price axes P^g. Thus

$$\hat{v} = v(b) \approx u_{0r}^{y\hat{r}}(p^{gr} + t^r, q^{gr})$$

for all $b \in B_{r1} \cup ... \cup B_{r5}$. This near-equality still leaves open the possibility of substitution between different price-quantity pairs at market gr. In Theorem 9, however,

we shall see that this cannot happen to any appreciable extent (as in p.h. equilibrium there will be at most two neighboring price levels being occupied at market gr).

Before stating Theorem 9 we shall introduce assumption (A11) of *bounded monotonicity of demanders' original utility functions* u_0^n: There exists $\tilde{\Delta} > 0$ such that for any $g \in G$, $n \in N^{gD}$

$$\tilde{\Delta} + \max_{q^g} u_0^n(b_{\neg p^g}, q^g) \geq \max_{q^g} u_0^n(b_{\neg p'^g}, q^g)$$

for all $b \in A^n$, whenever p'^g is just one discrete price step below p^g.

Notice: $\tilde{\Delta}$ in this assumption and in the bounded monotonicity assumption for original utility functions of *suppliers* may, without restriction of generality, be chosen identical, as in both cases they only serve as upper bounds.

The final assumption to be made is (A12) *continuity of the penalty functions* z^{ng} as functions of relative demand.

Theorem 9: Let $\hat{\underline{p}}$ denote some p.h. equilibrium state of the economy $(N,(A^n),(u^n(\cdot|\cdot)),(d^n(\cdot|\cdot)))$. Then under the previous assumptions (A1-12) the following holds:

1) At any market g at most two price levels in $\mathbf{P}^g$ are being occupied. If it is actually two, then the two price levels are neighboring ones.
2) If at market g two price levels are being occupied, πg^S and πg^D, say, (with $\pi g^S < \pi g^D$) then at level πg^S suppliers are not rationed, while at level πg^D demanders are not. In case suppliers are rationed at level πg^D or demanders at level πg^S, their rationing is negligible in the sense that

$$z^{ng}(r^g(\pi g^S)) < \tilde{\Delta} + 2\underset{\sim}{\Delta} \quad \text{for any } n \in N^{gD} \tag{14}$$

and

$$z^{ng}(r^g(\pi g^D)) < \tilde{\Delta} + 2\underset{\sim}{\Delta} \quad \text{for any } n \in N^{gS} \tag{15}$$

(with $\underset{\sim}{\Delta}$ as introduced in the strong monotonicity assumption for original utility functions of suppliers).

3) If at market g only one price level is being occupied, πg, say, then at market g suppliers may be non-negligibly rationed only when π^g is at the lowest price level in $\mathbf{P}^g$, i.e., $\pi^g = \min \mathbf{P}^g$, and demanders only when $\pi^g = \max \mathbf{P}^g$. More precisely,

$$z^{ng}(r^g(\pi g)) < \tilde{\Delta} + 2\underset{\sim}{\Delta} \tag{16}$$

for $\pi^g > \min \mathbf{P}^g$ and $n \in N^{gS}$ as well as for $\pi^g < \max \mathbf{P}^g$ and $n \in N^{gD}$.

(14), (15) and (16) may rightly be interpreted as negligible rationing for the reason that - under continuity of the original utility functions - $\tilde{\Delta}$ and $\underset{\sim}{\Delta}$ grow arbitrarily small as the mesh spacings of the various $\mathbf{P}^g$ become ever finer (i.e. as the maximal distance between any two neighboring price levels tends to zero), in certain cases only when the $\mathbf{Q}^g$ are being refined simultaneously.

Proof of Theorem 9: In this proof, $\tilde{P}$ will always designate some weak Nash equilibrium closely approximating $\hat{P}$ in the sense of (i) and (ii) above. Because of (iii) one will be allowed to assume $\tilde{P}$ to belong to any suitable (ρ,Δ)-class as may be required in the course of the proof.

Ad 1: Let us assume that at some market g there exist three price levels p'^g, p''^g, p'''^g such that $\hat{P}^S_{p'g} \geq \hat{P}^S_{p''g} \geq \hat{P}^S_{p'''g} > 0$ ($\hat{P}^S_{pg}$:= total number of suppliers at price level pg). As then for some suitable ρ'', $0 < \rho'' < 1$, not more than $\hat{P}^{gS}_{act}(1-\rho'')/(1+\rho'')$ of all suppliers active at market g find themselves concentrated on any two price levels, an analogous negative statement holds for $\tilde{P}$, with some ρ', $\rho'' < \rho' < 1$, in place of ρ''. This, however, contradicts Theorem 7. (Remember: $\tilde{P}$ may be assumed to belong to any $(\rho,\underset{\sim}{\Delta})$-class of $\underset{\sim}{u}(\cdot|\tilde{P})$, $\rho > 0$.) Therefore, in the p.h. equilibrium state $\hat{P}$, at any market g, at most two price levels are being occupied. If it is actually two, they are necessarily neighboring ones, because otherwise $\tilde{P}$ could not possibly exhibit two *neighboring* price levels, nor a single one carrying almost all of the suppliers active at market g. This, however, is requested by Theorem 7 for any state in any $(\rho,\underset{\sim}{\Delta}$-class of $\underset{\sim}{u}(\cdot|\tilde{P})$ with ρ sufficiently small.

Ad 2: If $\tilde{P}$ approximates $\hat{P}$ sufficiently closely, almost all of its active suppliers at market g are concentrated on price levels πg^S and πg^D. As, according to Theorem 7, the price levels pg^S and pg^D are the ones that carry almost all active suppliers [whenever $\tilde{P}$ belongs to some $(\rho,\underset{\sim}{\Delta})$-class of $\underset{\sim}{u}(\cdot|\tilde{P})$ with ρ sufficiently small], πg^S necessarily equals pg^S, and πg^D equals pg^D. In state $\tilde{P}$, at $\pi g^S = pg^S$, at least one population of demanders is rationed (by definition of pg^D) but no suppliers, and at $\pi g^D = pg^D$ all suppliers are rationed, but no demanders. That means, in state $\tilde{P}$ the inequalities $r^g(\pi g^S) > \alpha g > r^g(\pi g^D)$ hold (with αg as introduced in the complementarity of penalty functions assumption). By continuity of relative demand as a function of the occupation numbers, in state $\hat{P}$ the inequalities $r^g(\pi g^S) \geq \alpha g \geq r^g(\pi g^D)$ hold, which proves the first part of 2). - We shall show by contradiction that the inequalities (14), (15) also hold. Assume, e.g., that for some $n \in N g^D$ with $\hat{P}^n_{\pi g^S} > 0$ (14) does not hold. Then, in state $\tilde{P}$, $z^{ng}(r^g(\pi g^S)) \geq \tilde{\Delta} + \underset{\sim}{\Delta}$ (by way of continuity of the penalty functions z^{ng}) implying $\tilde{P}^n_{\pi g^S} \leq \rho\ \tilde{P}^n_{\pi g^D}$. As $\rho > 0$ may be assumed arbitrarily small [see (iii) above] the last inequality is not compatible with $\tilde{P}$ being an arbitrarily close approximation of $\hat{P}$; (notice $\hat{P}^n_{\pi g^S} > 0$). Thence, (14) holds. In an analogous way (15) is proven.

Ad 3: If, in state $\hat{P}$, $r^g(\pi g)$ happens to equal αg, none of the populations $n \in N$ are rationed at market g. In case $r^g(\pi g) \neq \alpha g$ consider some population n with $\hat{P}^n_{\pi g} > 0$ which is rationed at market g, i.e. $r^g(\pi g) > \alpha g$ in case $n \in N g^D$ or else $r^g(\pi g) < \alpha g$ in case $n \in N g^S$. In both cases the technique of proof is essentially the same and follows the scheme presented under 2): In a first step, from $z^{ng}(\pi g) > \tilde{\Delta} + 2\underset{\sim}{\Delta}$, to hold in state $\hat{P}$,

one deduces by way of continuity of the penalty functions z^{ng} that in state $\tilde{\underline{P}}$ $z^{ng}(\pi g) \geq \tilde{\Delta} + \underset{\sim}{\Delta}$. If then $n \in Ng^S$ and $\pi g > \min \mathbf{P}g$, one obtains $\tilde{P}^n_{\pi g} \leq \rho\, \tilde{P}^n_{\pi \underline{g}}$ (where $\pi \underline{g}$ designates the next lower price level in $\mathbf{P}g$). As $\rho > 0$ may be arbitrarily small the last inequality is not compatible with $\tilde{\underline{P}}$ being an arbitrarily close approximation to $\hat{\underline{P}}$. Thence, (16) holds for suppliers. As already said, the technique of proof is the same for demanders. This terminates the proof of Theorem 9.

The price levels πg^S and πg^D, or πg, respectively, appearing in Theorem 9 will be called *p.h. equilibrium prices*. When there are two p.h. equilibrium prices at some market g they are necessarily neighboring ones, for which reason we shall refer to definiteness of p.h. equilibrium prices.

Note: The assumptions (A1-12) made in proving definiteness of p.h. equilibrium prices comprise no restrictions concerning the dependence of the original utilities u^n_0 on quantities q^g. Note also that Theorem 9 does not preclude the possibility of having *empty markets* g, i.e. markets where there is no supply at any of the price levels p^g.

We now confront the extent to which p.h. equilibria possess Walrasian properties. In a first step one may ask whether, under p.h. equilibrium conditions, economic units are quantity optimizers - with respect to original (!) utility functions - for parametric price vectors. In this context it appears mandatory to reveal that reasonable penalty functions z^{ng} appear not only to depend on price levels (via relative demand), but also on quantities according to a law, e.g., of the type

$$z^{ng}(q^g; r^g(p^g(b))) = f^{ng}(q^g(b))\, \bar{z}^{ng}(r^g(p^g(b))),$$

where the f^{ng} are suitably chosen positive functions with non-negative slopes.

The results obtained so far continue to hold in this more general case, because the assumptions made referring to penalty functions, i.e. (A1), (A5), (A9), (A10), and (A12), may essentially be taken over unchanged (their verification in concrete situations not becoming more difficult than before). The main point is that the negligible rationing statements of Theorem 9 [inequalities (14), (15) and (16)] remain valid.

Therefore, it is of interest to note that for any p.h. equilibrium $\hat{\underline{P}}$, Theorem 8 in conjunction with the continuity argument in Remark 1 yields, for any $b \in A^n$ with $\hat{P}^n(b) > 0$, the equality

$$u^n_0(b) - \sum_{g \in Gb} z^{ng}(q^g(b); r^g(p^g(b))) =$$

$$\max_{a \in A^n(\hat{\underline{P}})} \{u^n_0(a) - \sum_{g \in Ga} z^{ng}(q^g(a); r^g(p^g(a)))\}, \qquad (17)$$

where $A^n(\hat{\underline{P}})$ consists of all those $a \in A^n$ for which $q^g(a) > 0$ occurs only if at the same time $p^g(a)$ is a p.h. equilibrium price pertaining to $\hat{\underline{P}}$, or more briefly: a is active solely at p.h. equilibrium price levels. (The reason is that only then is one assured that the relative demand values occurring on the r.h.s of (17) are all well defined.)

Equation (17) is a perturbed *quantity optimization* statement and implies the following inequalities:

For any $b \in A^n$ with $\hat{P}^n(b) > 0$

$$\max_{a \in A^n(\hat{\underline{P}})} u_0^n(a) \geq u_0^n(b) \geq \max_{a \in A^n(\hat{\underline{P}})} u_0^n(a) +$$

$$+ \sum_{g \in Gb} z^{ng}(q^g(b); r^g(p^g(b))) - \max_{a \in A^n(\hat{\underline{P}})} \sum_{g \in Ga} z^{ng}(q^g(a); r^g(p^g(a))), \quad (18)$$

(notice, $\hat{\underline{P}}^n(b) > 0$ implies $b \in A^n(\hat{\underline{P}})$). If $\hat{\underline{P}}$ is a negligibly rationed p.h. equilibrium (i.e. rationing at the p.h. equilibrium price level(s) of any nonempty market g is negligible in the sense specified in Theorem 9) then (18) yields

$$\max_{a \in A^n(\hat{\underline{P}})} u_0^n(a) \geq u_0^n(b) \geq \max_{a \in A^n(\hat{\underline{P}})} u_0^n(a) - |G|(\tilde{\Delta} + 2\underset{\sim}{\Delta}), \quad (19)$$

where $|G|$ is the cardinality of the set G of all commodities considered. For arbitrarily given $\varepsilon > 0$, the mesh spacings of the price axes P^g may be chosen to be so fine that $|G|(\tilde{\Delta} + 2\underset{\sim}{\Delta}) < \varepsilon$ holds (under suitable continuity conditions for the original utility functions u_0^n, and possibly only after a suitable refinement of the mesh spacings of the quantity axes, too). We shall express this more briefly by saying that in case the p.h. equilibrium $\hat{\underline{P}}$ is only negligibly rationed, any alternative b, of some population n, that is actually being chosen in state $\hat{\underline{P}}$ is *almost optimal.*

Remark 5: In Remark 4 we started to discuss spatial substitution in regard to the example of multizonal shopping. Let us now assume the following very simple type of original utility function:

$$u_{0r}^{y\hat{r}}(p^{gr} + t^r, q^{gr}) := u_{00}^{y\hat{r}}(q^{gr}) - d_0^{\infty} \quad (20)$$

where d_0^{∞} equals ∞ or 0 depending on whether the budget constraint $(p^{gr} + t^r)q^{gr} \leq y$ is violated or not. The fact that the functions $u_{0r}^{y\hat{r}}$ all coincide with a single function $u_{00}^{y\hat{r}}$ is to express that the various shopping zones all offer the same composite good. As common in neoclassical microeconomic theory, we assume $u_{00}^{y\hat{r}}$ to be a strictly increasing function of q^g. We proceed to consider some negligibly rationed p.h. equilibrium $\hat{\underline{P}}$, and make the following natural assumptions:

(i) There is at least one shopping zone r with its market gr nonempty.

(ii) The income y of population $y\hat{r}$ is sufficiently high in order to allow for shopping to occur in at least one shopping zone, i.e. there exists $b \in A^{y\hat{r}}$ such that $\hat{P}^{y\hat{r}}(b) > 0$.

For $b \in A^{y\hat{r}}$, active at some market gr, let r(b) denote that shopping zone where shopping takes place when alternative b is being exercised. Let p(b) and q(b) then denote

related price and quantity levels, respectively. Furthermore, for any $b \in A^{y\hat{r}}$ with $q(b) > 0$, let bb denote that alternative $A^{y\hat{r}}$, active at the same market as b at the same price level as b, the quantity component of which, however, is maximal without violating the budget constraint $[p(b)+t^{r(b)}]\, q(bb) \leq y$. As a consequence of the particular type (20) of original utility functions we have adopted here, the following inequality holds:

$$u_0^{y\hat{r}}(bb) \geq u_0^{y\hat{r}}(b). \tag{21}$$

Thus, as for $\hat{p}^{y\hat{r}}(b) > 0$, b is almost optimal (cf. 19)), bb is almost optimal, too. - We are now going to show that any two almost optimal alternatives b and c have quantity components that are almost equal (in a sense to be specified presently by some inequality). Our point of departure is inequalities (19), with $A^n(\hat{\underline{p}})$ consisting of all those alternatives $b \in A^{y\hat{r}}$ that do not involve shopping activity at any non-p.h. equilibrium price level ($A^n(\hat{\underline{p}})$ containing thus, e.g., the no shopping at all alternative). As (19) holds for b, and also for c when inserted in place of b,

$$|u_0^{y\hat{r}}(b) - u_0^{y\hat{r}}(c)| \leq |G|(\tilde{\Delta} + 2\underset{\sim}{\Delta}), \text{ i.e.}$$

$$|u_{00}^{y\hat{r}}(q(b)) - u_{00}^{y\hat{r}}(q(c))| \leq |G|(\tilde{\Delta} + 2\underset{\sim}{\Delta}).$$

If $\ell > 0$ is a lower bound to average slopes of $u_{00}^{y\hat{r}}$, then the last inequality yields

$$|q(b) - q(c)| \leq \ell^{-1} |G|(\tilde{\Delta} + 2\underset{\sim}{\Delta}).$$

This is the above-mentioned inequality specifying what is meant by "$q(b) \approx q(c)$". Notice, as ℓ and G are independent of the mesh spacings chosen for price and quantity axes, the difference between q(b) and q(c) is made arbitrarily small by rendering $\tilde{\Delta} + 2\underset{\sim}{\Delta}$ sufficiently small. - To sum up the arguments of this Remark: If $\hat{\underline{p}}$ is a negligibly rationed p.h. equilibrium and $\hat{p}^{y\hat{r}}(b) > 0$, $\hat{p}^{y\hat{r}}(c) > 0$, then $q(b) \approx q(c)$, whether $r(b) = r(c)$ or not: individual demand by economic agents from $y\hat{r}$ is almost the same in every shopping zone where shopping by population $y\hat{r}$ actually occurs.

Remark 6: Aggregate (satisfied) demand, in some arbitrary shopping zone r', by population $y\hat{r}$, is almost equal to $q(b)\hat{p}_{r'}^{y\hat{r}}$ if b is any alternative with $r(b) = r'$, $\hat{p}^{y\hat{r}}(b) > 0$, and if $\hat{p}_{r'}^{y\hat{r}} := \sum \hat{p}^{y\hat{r}}(a)$ where summation extends over all alternatives $a \in A^{y\hat{r}}$ with $r(a) = r'$ (and $\hat{p}^{y\hat{r}}(a) > 0$). Notice, $\hat{p}_{r'}^{y\hat{r}}$ is only known after $\hat{\underline{p}}$ has been calculated. This is not necessarily alike for *overall* aggregate demand by propulation $y\hat{r}$, i.e. for

$$\sum_{r'} q(b)\, \hat{p}_{r'}^{y\hat{r}} = q(b) \sum_{r'} \hat{p}_{r'}^{y\hat{r}}.$$

If the no shopping at all alternative b_0 has low original utility (which is normal to posit), then $\hat{p}^{y\hat{r}}(b_0) = 0$, and, thus

$$\sum_{r'} \hat{p}^{y\hat{r}}_{r'} = \hat{p}^{y\hat{r}},$$

$\hat{p}^{y\hat{r}}$ designating the total number of economic agents in population $y\hat{r}$. In some models $\hat{p}^{y\hat{r}}$ is dealt with as an exogenous parameter and, thus, known from the outset. If, however, in some bilevel model of an economy choice of location is part of the modelling exercise (see the end of Section 2), then $\hat{p}^{y\hat{r}}$ is an endogenous variable and only known after determination of the p.h. equilibrium $\hat{p}$.

Remark 7: As a final point in connection with the preceding simple shopping model - treated here as part of a more comprehensive general equilibrium model - we shall discuss full price incurred by population $y\hat{r}$ when shopping in zone r'. We already know that in any negligibly rationed p.h. equilibrium $\hat{p}$, for any $b \in A^{y\hat{r}}$ with $\hat{p}^{y\hat{r}}(b) > 0$, $q(b)$ and $q(bb)$ are almost equal (see Remark 5). In addition, for sufficiently fine mesh spacings of the quantity axes, Q^g, $q(bb)[p(b) + t^{r(b)}] \approx y$. By use of $q(b) \approx q(bb)$, one deduces $q(b)[p(b) + t^{r(b)}] \approx y$. A similar approximate equality holds for any other alternative a with $\hat{p}^{y\hat{r}}(a) > 0$: $q(a)[p(a) + t^{r(a)}] \approx y$. As $q(a) \approx q(b)$, one finally obtains $p(b) + t^{r(b)} \approx p(a) + t^{r(a)}$. Thus, with the particular type (20) of original utility functions, $u_{0r}^{y\hat{r}}$, the full cost incurred when economic agents from $y\hat{r}$ shop in different zones is almost the same everywhere. - Unit transport cost, t^r, by the way, is not necessarily fixed and may be an endogenous variable.

Besides quantity optimization with respect to original utility, *market clearing* is a second Walrasian postulate to discuss. The natural definition of market clearing appears to be equality of current supply, on one hand, and the sum of purchases (i.e. total demand in terms of the above developed terminology) and depreciation (in physical units) of stock, on the other. The latter term only comes to bear when the commodity considered is storable. In that case market clearing is a stationarity requirement because the stock of a storable good, the market of which is cleared in the preceding sense, has a vanishing time derivative. Thus, if one wants equilibrium to exhibit market clearing for storable commodities, the dynamics of the economic system modelled ought to be such that equilibrium states are obtained as stationary states. This touches upon the topic of passiveness of the system dynamics, briefly discussed by Dejon and Graef (1983), by Güldner (1984), by Dejon (1986), and in a more abstract setting by Wenzel (1982). We shall not elaborate on this point.

For labor and services (nonstorable commodities with permanent zero stock), in direct and in p.h. equilibrium, overall current supply (totalled over all price levels) is typically larger than overall purchases. The quotient of overall purchases by overall supply at any single market g of labor or services is so-called average relative demand $\bar{r}^g$, at that market, and may be obtained as some weighted average of the relative demand values $r^g(p^g)$ at the various price levels p^g. In case of p.h. equilibrium, at the p.h. equilibrium price levels relative demand is almost equal to α^g (see Assumption (A5) and Theorem 9) and, therefore $\bar{r}^g \approx \alpha^g$. It appears expedient to declare equilibrium state markets to be cleared

markets, thus introducing implicitly a 'natural' rate of unemployment of labor or service resources, respectively. The fact that the value of equilibrium average relative demand, $\bar{r}^g$, depends rather strongly on the choice of disutility functions z^{ng} (in the p.h. limit it is essentially only α^g that matters) underlines once more the importance of these functions.

5. CONCLUSIONS AND EXTENSIONS

Because of its conceptual structure, direct equilibrium modelling is comparatively easy to implement by computer. One of the helpful features is a rather extensive modularity, which one expects to detect by simply examining the listing of the main constitutive elements of any weak Nash equilibrium system, $(N, (A^n), (u^n(\cdot|\cdot)), (d^n(\cdot|\cdot)))$. However, one of the major drawbacks of direct equilibrium modelling lies in the large number of unknowns. To each market g, e.g., there are $|P^g| \cdot |Q^g|$ unknowns ($|P^g|$ and $|Q^g|$ denoting the cardinalities of P^g and Q^g, respectively). Yet, to add another positive feature, there is no need to devise market supply or demand functions. Instead, naturally, one has to set up (strict) utility functions. A pertinent characteristic feature is that market clearing - even in equilibria delicate enough to define at markets of nonstorable commodities like labor, capital services, and services in general (cf. end of Section 4) - is only introduced indirectly by way of penalty functions, which can be interpreted as search costs on the demand side of markets and as costs of uncertainty of demand on the supply side.

Weak Nash equilibrium problems are fixed point problems by their very definition. One may solve them iteratively, yet on the assumption that the economic system to be modelled is frequently enough out of equilibrium it appears preferable to try to emulate not only economic equilibrium states, but also disequilibrium time paths of the system - with the obvious prerequisite of adequate dynamic laws. Quite a broad class of such dynamic laws has been described by Dejon (1983) as 'attraction-regulated dynamic equilibrium' laws. These provide for a general setting of ideas which have been expounded in more specific contexts by, e.g., Allen and Sanglier (1979) and Dejon and Graef (1983).

Güldner (1984) described the first exploratory simulation experiments utilizing direct equilibrium modelling of a highly aggregated nonspatial economy, with a focus on aggregate output, purchases, price and wage inflation as well as rate of unemployment. He also reported about simulation in connection with location and land use modelling. Here the idea was to check whether direct equilibrium modelling would lead to the type of results expected on the basis of Alonso's (1964) analyses. The outcome was positive, confirming, e.g., numerically that the economic activity with the steepest bid rent curve locates closest to the CBD. Note that the related theory refers to p.h. equilibrium while the numerical calculations are for approximating direct equilibria which still exhibit the phenomenon under scrutiny. Dejon (1986) contains a very brief account of a simulation experiment conducted by H. Körner, which was designed to study the reaction of a direct equilibrium model to a change of the parameter α^g (see assumption (A5)) at one of the commodity markets.

Existence proofs for weak Nash equilibria, based on Brouwer's fixed point theorem, have been provided in particular applications; see, e.g., Müller (1980) for the traffic assignment problem. For direct equilibria, the existence proof appears to be less trivial as relative demand no longer depends continuously on the state vector $\underline{p}$ whenever $\underline{p}$ implies

vanishing supply at some price level (cf. Remark 1). For the time being, one simply trusts in the existence of direct equilibria, as well as in their approximate computability, by tracing time paths of systems under study until these have ostensibly settled down to an (approximate) stationary state. Until now, numerous such time paths have been calculated and, under reasonable choices of parameters, have always been observed to settle down. There remains, however, more analytical work to be done on asymptotic stability as well as on conditions of uniqueness of direct equilibria.

REFERENCES

Allen, P.M. and M. Sanglier, 1979, "A Dynamic Model of Growth in a Central Place System", *Geographical Analysis*, 11:256-272.

Alonso, W., 1964, *Location and Land Use*, Harvard University Press, Cambridge, Mass.

Dafermos, S.C., 1971, "An Extended Traffic Assignment Model with Applications to Two Way Traffic", *Transportation Science*, 5:366-389.

Daganzo, C.F. and Y. Sheffi, 1977, "On Stochastic Models of Traffic Assignment", *Transportation Science*, 11:253-274.

Dejon, B., 1983, "Attraction-regulated Dynamic Equilibrium Models of Migration of the Multiplicative Type", in D.A. Griffith and A.C. Lea, (eds.), *Evolving Geographical Structures*, NATO ASI Series, Proceedings of the ASI at San Miniato, Italy, July 1982.

Dejon, B., 1986, "Modelling an Economy in Space and Time: the Direct Equilibrium Approach with Attraction-regulated Dynamics", in D.A. Griffith and R. Haining, (eds.), NATO ASI Series, Proceedings of the ASI at Hanstholm, Denmark, August 1985.

Dejon, B. and F. Graef, 1983, "Eine Klasse von Modellen zur Beschreibung von städtischen Agglomerations- und Deglomerationsprozessen", in *Stadt, Region, Land*, Institut für Stadtbauwesen, Rheinisch-Westf. TH Aachen, 59:32-38.

Dejon, B. and B. Güldner, 1985, "Dynamics of Wide-sense Migrational Systems: a Choice Theoretic Approach", *Papers of the Regional Science Association*, 55:121-133.

Devarajan, S., 1981, "A Note on Network Equilibrium and Non-cooperative Games", *Transportation Research*, 15B:421-426.

Evans, S., 1973, "A Relationship Between the Gravity Model for Trip Distribution and the Transportation Problem in Linear Programming", *Transportation Research*, 7:39-61.

Güldner, B., 1984, "Attraktivitätsgesteuerte dynamische Alternativenwahlmodelle: Analysen und Simulationen in drei Anwendungsbereichen", Ph.D. Dissertation, Institute for Applied Mathematics, University of Erlangen-Nürnberg, Erlangen.

Haurie, A. and P. Marcotte, 1985, "On the Relationship between Nash-Cournot and Wardrop Equilibria", *Networks*, 15:295-308.

Iri, M., 1969, *Network Flow, Transportation and Scheduling*, Academic Press, London, New York.

McFadden, D., 1973, "Conditional Logit Analysis and Qualitative Choice Behavior", in P. Zarembka, (ed.), *Frontiers in Econometrics*, Academic Press, New York.

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

Müller, R., 1980, "Bestimmung optimaler Parameter bei Wegewahlmodellen vom Gleichgewichtstyp mit Gradientenverfahren", Ph.D. Dissertation, Institute for Applied Mathematics, University of Erlangen-Nürnberg, Erlangen.

Wenzel, G., 1982, "Zur Existenz und Stabilität von Lösungen gewisser impliziter Differentialinklusionen zur Beschreibung des dynamischen Gleichgewichts in abstrakten Netzwerken", Ph.D. Dissertation, Institute for Applied Mathematics, University of Erlangen-Nürnberg, Erlangen.

Advances in Spatial Theory and Dynamics
A.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 8

Rivalrous Consonance: A Theory of Mature Oligopolistic Behavior in a General Equilibrium Framework

R.E. Kuenne

1. INTRODUCTION

The theory of rivalrous consonance as a framework within which to analyze the strategies of mature oligopolistic industries incorporates a body of assumptions that appear to me to be widely accepted. Indeed, they receive such common acceptance as to seem trite. Yet, in my view, they have not previously been employed as the formal, postulational basis for an operational theory of oligopolistic decision making. The following are among the most important.

Assumption 1. Mature oligopolies, or rivalrous industries with a substantial industrial history, are *communities* in important respects. Individual units within such communities have important competitive interests which make them rivals in their goal seeking. But as members of an acknowledged community they have common interests which imply cooperative relations. Their actions, therefore, will be motivated by a blend of rivalrous and cooperative goals in mutual recognition of a *rivalrous consonance of long-run interests.*

Assumption 2. Each such community, in which individual actors are few enough in number to impact the industry in personally identifiable ways, has a power structure, or a web of perceptions among firms that has an important bearing upon their decisions. The binary, firm-to-firm, combination of rivalry and cooperation that constitutes an important component of their decision making is the operational expression of that power structure. Unless that sociological matrix of power relationships is incorporated in the analysis of industry decision making little hope exists for useful insights.

Assumption 3. Because these industrial communities and their power structures are the result of unique historical evolution energized by unique individuals and framed by unique industry and product characteristics, each is marked by distinguishing patterns of behaviour. A universal theory of oligopoly is therefore unattainable. Analytical ambitions must be limited to studies of industries *sui generis* with the goal of gaining insights into their functioning and structure, and with cautious generalizations arising from limited commonalities such industries may reveal.

Assumption 4. Corporations consistently reveal risk aversion in their attitudes to uncertain events, and this must project into the conditions of decision stability in their industries. The role of the cooperative motivation in tempering the rivalrous drive is thereby strengthened to the extent the latter threatens those stability conditions. Indeed,

one of the social functions of the power structure is to exercise such restraining influences.

Assumption 5. Corporations are multiobjective entities making decisions in a multidimensional target variable space incorporating price, quality, and advertising dimensions. They consist of sub-bureaucracies with their selfish goals, many of which clash with like goals of their fellow groups. Corporate policy, therefore, reflects these goals with different priorities, and is frequently ill-defined or even inconsistent. The goals, therefore, are distinct to the firm, and must be isolated to approximate the firm's motivation in a formal model.

Assumption 6. Few market structures are characterized by pure competition or monopoly. Most mixtures of competition and monopoly reveal themselves to be clusters of oligopoly. Therefore, if general equilibrium theory is to approach a realistic and operational form it must incorporate in an extensive manner oligopolistic decision making. By Assumption 3, such incorporations must have limited aims for generality. The extensive use of mathematical theorem deduction must give place to simulation analyses with numerical structure. The goal must no longer be the derivation of universal theorems from mathematically manipulable but pitifully unrealistic models, but of insights into the functioning of models provided with parametric scenarios of efficient design to yield useful insights.

If these assumptions approach the realistic, frameworks must emerge to encompass them. But oligopoly theory today is dominated by game theory, which only partially and rather imperfectly captures the essence of this environment. Its primary focus is upon the rivalrous implications of oligopolistic interdependence, even when it deals with cooperative aspects of such relationships. It is a single-objective analysis; it neglects industrial power structures as well as other sociological aspects of the decision making, and is not well suited to general equilibrium analysis.

The rivalrous consonance framework is an attempt to move analysis in the indicated directions and, unfortunately, given the almost exclusive attachment of present analysis to it, away from game theory in vital ways. It is experimental, but at least it is designed for application to realistic oligopolies and is therefore testable. I shall present it in its most rudimentary form in the present paper, but even at its most complicated it is a "middle-brow" theory in terms of complexity of structure.

2. THE OLIGOPOLY IN RIVALROUS CONSONANCE

Assume an oligopoly with n rivals producing differentiated brands in a product group. Suppose their strategy consists of setting price with the primary goal of maximizing profits subject to (1) industry ties to rivals, (2) groups of subsidiary goals unique to each firm, and (3) demand and cost functions. If we focus upon rival i's motivation, rivalrous consonance views its objective function as an "extended profit function", E_i:

$$\underset{p_i}{\text{Max}}\ E_i = \sum_{j=1}^{n} \theta_{ij}(p_j - k_j)x_j\,, \tag{1}$$

subject to a set of constraints, where

p_j = price of rival j's product

$x_j = a_j - b_{jj}\,p_j + \sum_{j \neq k} b_{jk}\,p_k + B_{jq}\,Q$ = rival j's demand function

Q = a vector of unspecified variables (other than industry prices) affecting demand

k_j = average cost, assumed constant, of product j

Θ_{ij} = binary consonance factors for firm i relevant to rivals j, $\Theta_{ii} \equiv 1$.

The distinctive terms in the objective function are elements in the binary consonance matrix, Θ, which specifies the effective power structure of the industry. Rival i maximizes his profits and the profits of his rivals when those rival profits are discounted by Θ factors. $\Theta_{ij} \in \Theta$ is the dollar amount rival i values a \$1 profit or loss for rival j. Thus, if $\Theta_{ij} = \$.25$ a dollar change in rival j's profit is the equivalent of \$.25 in rival i's own profit. The ith row of Θ defines rival i's attitudes to each of its competitors as those attitudes impact on its decision making. The ith column of Θ, of course, defines rivals' power structure attitudes to rival i. Hence, rival i's perception of rival j's ability to retaliate against it should rival i's price decisions impact too severely rival j's profits is included in rival i's decision making. But it also includes rival i's concern for rival j's welfare on more altruistic grounds as well, perhaps with some concern for the peaceful stability of the industry. Of course Θ_{ij} need not equal Θ_{ji}, since rivals' perceptions of each other's position in the power structure can be quite disparate.

The subsidiary goals of rival i are incorporated into the model as constraints. To illustrate, let us use a set of common goals exhibited by firms in oligopoly.

First, output must be within the firm's capacity to produce, q_i:

$$C^1 : x_i - q_i \leq 0. \tag{2}$$

Second, the firm insists upon attaining a minimum level of sales, m_i, set perhaps by an expected target share proportion:

$$C^2 : m_i - x_i \leq 0. \tag{3}$$

Third, the firm sets upper and lower bounds, t_i^+ and t_i^- respectively, on price changes in the period, specified as proportions of previous period price $\bar{p}_i$:

$$C^3 : p_i/\bar{p}_i - t_i^+ \leq 0, \tag{4}$$

$$C^4 : t_i^- - p_i/\bar{p}_i \leq 0. \tag{5}$$

Lastly, price is bounded away from a specific lower limit which we will assume to be zero but may not be:

$$C^5 : u_i - p_i \leq 0. \tag{6}$$

The rivalrous consonance framework includes such "crippled optimization" submodels that consist of nonlinear programming models. Each is solved sequentially assuming all rivals' prices are temporarily fixed so that if each rival submodel is convex in own price the sufficient condition for a global constrained maximum will be achieved. Since all constraints are linear they are convex. The objective function yields

$$d^2E_i/dp_i^2 = -2b_{ii} < 0, \tag{7}$$

and hence it is (strictly) concave in p_i. Hence, each submodel is a convex nonlinear programming model.

It should be noted that such firm models may not be convex and this must be accepted as a possible reflection of the real oligopolistic world. Local maxima may therefore be all that can be attained realistically. Convexity is convenient, to be sure, but is has no claim upon exclusive consideration by the theorist. I have solved rivalrous consonance models which were not convex for 5 and 11 firms using the Fiacco/McCormick Sequential Unconstrained Minimization Technique (SUMT) with no problems.

For this paper I will remain with the simple profit maximization model limited to one industry given in (1)-(6) above. However, let me indicate the flexibility and extendability of the approach briefly. First, different objective functions may be adopted to accommodate firms' differing goal sets. For example, a firm's major objective may be to attain a target rate of return, or to maximize market share, and different rivals in the same industry may have different major objectives as well as subsidiary objectives. This causes no problem to the modelling. Price leadership and followership are easily included for relevant firms as major objectives.

Second, general equilibrium is attained by linking an industry to its suppliers and its customers. Rival i's industrial customers as well as its rivals are given consonance factors and included in the extended profit function. In this manner patterns of price behavior among client industries can be studied given parametric shocks (see Kuenne, 1986, for an extended presentation).

3. A GRAPHIC PRESENTATION

In order to depict some important characteristics of the pricing let us simplify by assuming a duopoly under rivalrous consonance and let us ignore constraints. Then reaction function depictions of some limiting cases are possible that serve to unify certain aspects of classic oligopoly theory by revealing them to be limiting cases of rivalrous consonance.

Figure 1 illustrates some basic concepts and implications of rivalrous consonance for the case of duopoly.

For rival i, first order necessary conditions for a crippled optimization price solution, given the price for rival j, p_j, are

$$\partial E_i/\partial p_i = (x_i - b_{ii}(p_i - k_i)) + \theta_{ij}\, b_{ij}(p_j - k_j),\ \ i = 1,2;\ \ j \neq i, \tag{8}$$

and, solving the pair of equations for p_1 and p_2, we obtain

$$p_i^{\theta} = \frac{-(2b_{jj}M_i + (b_{ij} + \theta_{ij}b_{ji})M_j)}{4b_{ii}b_{jj} - (b_{ij} + \theta_{ij}b_{ji})(b_{ji} + \theta_{ji}b_{ij})} = \frac{N_i^{\theta}}{D^{\theta}},\ \ i = 1,2;\ \ j \neq i, \tag{9}$$

where $M_i = \theta_{ij}b_{ji}k_j - a_i - b_{ii}k_i$. Solving in terms of price interactions, we get

$$p_i^{\theta} = (0.5/b_{ii})[a_i + b_{ii}k_i - \theta_{ij}b_{ji}k_j + (b_{ij} + \theta_{ij}b_{ji})p_j],\ \ \ i = 1,2;\ \ j \neq i \tag{10}$$

These are the reaction functions for the two firms given the consonance factor vector $\theta = [\theta_{ij}, \theta_{ji}]$. Their values yield $p^{\theta} = [p_i^{\theta}, p_j^{\theta}]$, the joint extended profit equilibrium.

For the mature oligopoly we hypothesize that $\theta_{ij} \in [0,1]$ and $\theta_{ji} \in [0,1]$. If a θ-value becomes negative the rival adopting it is waging price war, valuing his rival's loss at a positive own profit value. At $\theta_{ij} = 1$, rival i values rival j's profit as equal to his own. It is difficult to imagine conditions in which he might value his rival's profit at a greater value

than his own. Hence the lower and upper bounds adopted, although no conceptual problem arises if the consonance factors are allowed to broach these limits.

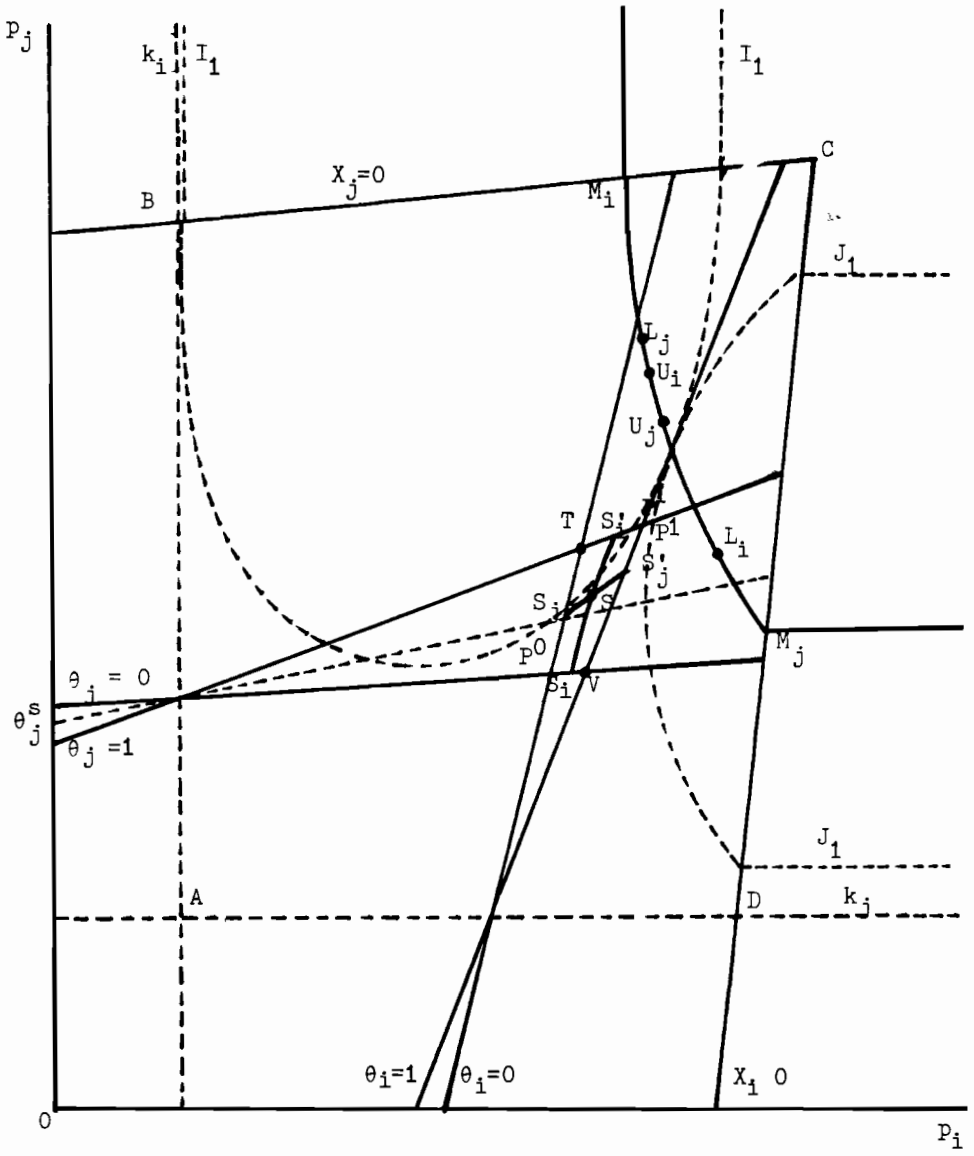


Figure 1 Isoprofit contours, reaction functions, and solution regions in duopoly

If $\theta_{ij} = \theta_{ji} = 0$, we have the case where each rival makes his pricing decision in myopic disregard of its impact upon his rival. I have called the resulting P^0 vector the *Cournot point*, in the extended sense that each rival ignores the impact of his actions upon his competitor's objective function. The fact that Cournot used quantity instead of price as the firm's target variables is unimportant. The truly interesting fact is that we can

reinterpret Cournot behavior to be equivalent to ignoring the impacts of one's decisions upon the welfare of one's competitor. This behavior is a more credible interpretation than the standard one of assuming one's rival's price (or quantity) will remain unchanged as one alters one's own.

If we go to the other extreme and assume $\theta_{ij} = \theta_{ji} = 1$, we have P^1 which I have called the *Chamberlin point*, in honor of that economist's suggestion that mutual interdependence recognized leads to a joint profit maximization. Rivalrous consonance also permits a reinterpretation of Chamberlin behavior: it is not mutual interdependence recognized that distinguishes this limiting case from the Cournot case. System (8) makes it quite clear that both cases force the rival's parameters into the decision making of the firms. The true distinction is the degree of valuation placed by each firm upon its rival's profit welfare. In this sense, rivalrous consonance is at once a unifier of classic oligopoly theory and an extension of it.

Figure 1 graphs the price space for duopoly and the P^0 and P^1 limiting solutions. The dashed lines depict the constant average costs, k_i and k_j, of the rivals. The straight lines that start at the axes and terminate at C are the loci of price vectors at which x_i and x_j become zero. Hence, the relevant price space for decisions is found in the polygon ABCD, since we assume that neither firm will operate in the long-run where profits are negative.

Isoprofit functions for the firms, of which we have drawn only two, I_1 and J_1, are ellipses in the relevant region and become horizontal or vertical lines beyond that region. Consider now the straight line reaction functions (defined in (8)) when $\theta_{ij} = \theta_{ji} = 0$, labelled θ_i and θ_j to simplify notation in the figure. These intersect the firms' isoprofit contours at their minimum points, i.e., where they are tangent to horizontal or vertical lines for firm i and j respectively. Their intersection at P^0 yields the Cournot solution. At the other extreme, consider the reaction functions where $\theta_{ij} = \theta_{ji} = 1$: if we draw the joint profit isocontours (which we have not in Figure 1) these reactions functions would intersect them where dp_j/dp_i or dp_i/dp_j are zero. The intersection of these reaction functions yields P^1, the joint profit maximum.

Consider the set of points M_iM_j on which P^1 lies. M_i is the monopoly profit point for rival i, or the P at which rival j is eliminated from the market and rival i is maximizing its profit. It is the point upon which successively higher I isoprofit contours converge. M_j is the monopoly point for rival j with similar interpretation. The line M_iM_j is the locus of tangencies of I and J contours, or all points such that rival i cannot be bettered in profit receipts without harming rival j, and vice versa. P^1 lies on this locus at the point where the sum of such tangent contours is a maximum. If collusion were allowed, we would expect a negotiated bargain to be struck somewhere on this "negotiation set" between the exclusive beneficiary limiting points M_i and M_j. We will narrow the expected limits of negiotiated solutions further below.

The important potential of rivalrous consonance, however, lies in its ability to permit θ-factors to assume intermediate values between 0 and 1, and thereby depict the pricing results of less extreme power structures. In Figure 1 the dashed reaction function for $0 \leq \theta_j^s \leq 1$ has been drawn to illustrate the new flexibility. Where collusion is ruled out of consideration, so the negotiation set (other than P^1) can be ignored, we can set tighter bounds on the feasible solution region. When rival i acts myopically and rival j altruistically the P-solution occurs at T, and when roles are reversed at V. Hence, the region P^0TP^1V is the region of feasible solutions for the mature oligopoly as we have interpreted it.

Let us now suppose that in the very long-run the consonance factors are variable and will come to reflect more accurately the egoistic profit drives of the rivals tempered by their competitors' power and acumen in the competitive process. One of the important

desires of each opponent will be to assess the Θ-factor that governs his competitor's strategy and to use such knowledge to improve his profit position. For example, let us start with $\Theta = [0,1]$ with consequent P = T. Suppose rival j surmises that rival i has $\Theta_i = 0$ and that it will for long periods retain that perception of its role in the industry power structure. Then it will pay rival j to find his *Stackelberg point* on $\Theta_i = 0$ where he reaches the innermost J contour he can attain - say at S_j where his effective $\Theta_j = \Theta_j^S$. That J profit contour is the minimum profit that rival j need accept as long as rival i adopts Θ_i, and, since $\Theta_i = 0$ is by our assumption the lowest value rival i can adopt, this is the minimum profit rival j need accept under any condition . Hence, if we move along that J contour to its intersection with the M_iM_j negotiation set (say at L_j) this places a lower bound on the negotiated distribution of joint profit to rival j. Also, if the T contour that goes through S_j is followed to its intersection with M_iM_j (say at U_i) we have the maximum amount of joint profit that rival i can obtain by negotiation.

Symmetrically, if we start with $\Theta_i = 1$ and $\Theta_j = 0$, and P = V, and rival i seeks its Stackelberg point (say at S_i), we can locate L_i and U_j on the negotiation set. It would then benefit both parties to arrive at a P on M_iM_j between U_i and U_j by collusion, and were such action admissible this segment of M_iM_j would hold strong interest for us. But we have ruled out the possibility of collusion and must return to the rivals' Stackelberg points.

Arbitrarily, suppose that is S_j. If rival i surmises that rival j will move along the reaction function Θ_j^S, it will find that point on Θ_j^S that is its new Stackelberg solution, different from S_i. That implies a new Θ_i^S (not drawn) and rival j will use it to obtain yet another Stackelberg point. If we plot the loci of each rival's Stackelberg points we obtain their *Stackelberg functions*, plotted on Figure 1 as S_jS_j' and S_iS_i'. These may be viewed as their long run reaction functions, and their intersection at the point S I have termed the generalized Stackelberg point. It is a Nash equilibrium.

The point S within the feasible solution region P^0TP^1V has a strong appeal because from any point in the subset $P^0S_jSS_i$ it is possible for both firms to improve their profits by moves toward S, until S is achieved. For any P in the subset $SS_i'P^1s_j'$ it will always benefit one of the noncolluding rivals to move toward S at the expense of the other - a one-sided benefit game that ends at S. Hence, in the very long run, when power structures alter under the motivating force of profit maximization, we should expect P to lie in the subset $SS_i'P^1s_j'$ with S as the point of gravitational pull.

In the short run, however, Θ is fixed and we return to the short-run reaction functions.

4. COMPARATIVE STATICS PROPOSITIONS

Trivially, $dp_i/d\Theta_i > 0$ if $p_i \leq k_i$ and $dp_i/d\Theta_j > 0$ iff $p_i > k_i$. Consider now the impact of cost increases with fixed Θ. Let us rewrite the reaction function (10) as

$$p_i^{\Theta} = A_i + B_i p_j^{\Theta}, \tag{11}$$

where $A_i = 0.5(a_i + b_{ii}k_i - \Theta_{ij}b_{ji}k_j)/b_{ii}$ and $B_i = 0.5(b_{ij} + \Theta_{ij}b_{ji})/b_i$. Then, for the intercept terms

1. $dA_i = 0.5dk_i$
2. $dA_j = -(0.5b_{ij}\theta_{ji}/b_{jj})dk_i.$ (12)

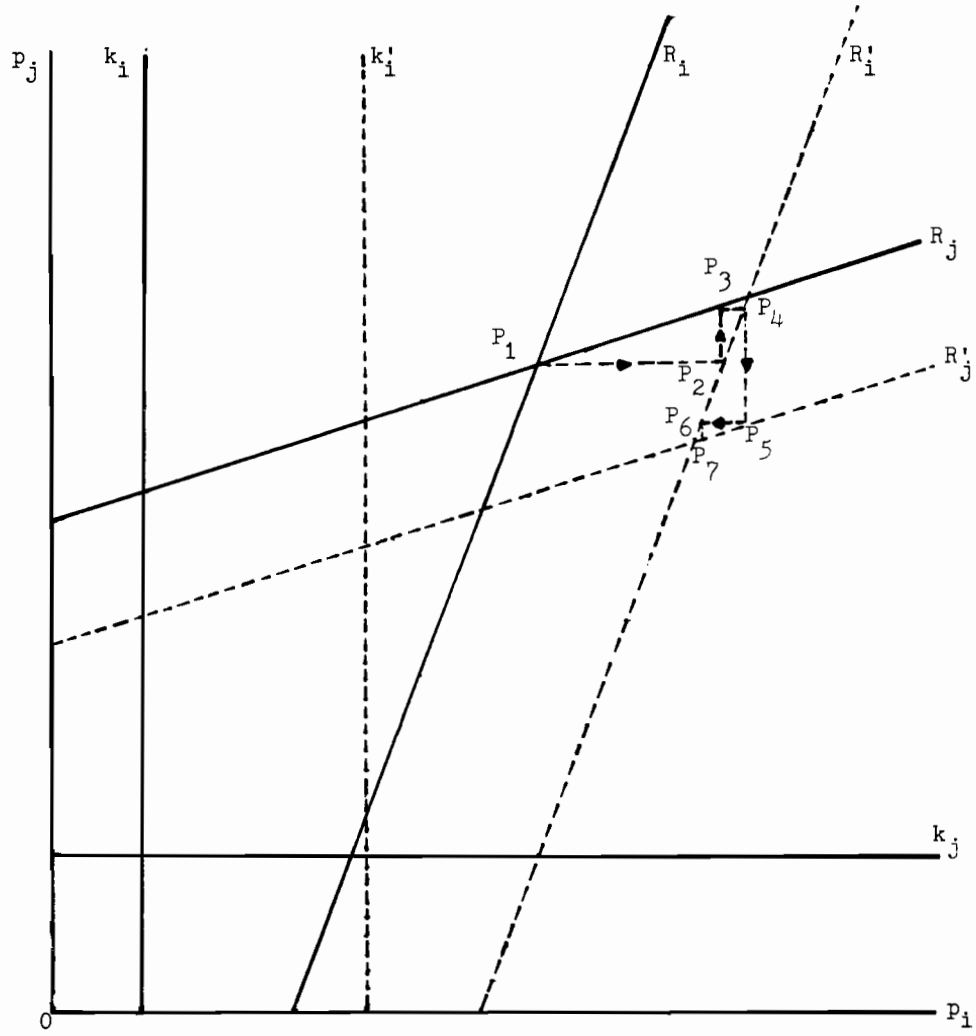


Figure 2 Comparative statics of a cost change

For an own-cost change the reaction function R_i on Figure 2 shifts to the right by one half the rise in cost; however, R_j shifts down, reducing the rise in p_i that would occur if R_j remained fixed. The adjustments are shown in Figure 2, and it will be useful to work through the adjustment process in achieving the new equilibrium.

R_i shifts by one half the cost increase to R_i' and firm i, assuming p_j constant at p_j^1 moves the price vector to P_2 from P_1. For the moment R_j remains fixed, and with p_i changed induces firm j to move to P_3, further inducing firm i to move the solution to P_4.

Obviously, the amount of price movement to this point depends upon the slope of R_j. The rise is greater the less sensitive x_j is to p_j (i.e. the smaller is b_{jj}), the more sensitive it is to p_i (i.e., the larger is b_{ji}), and the larger is Θ_{ji}. Rivalrous consonance boosts prices, holding an umbrella over rival i's head. Phase I ends with P_4, which implies p_i and p_j higher than in P_1.

At the start of Phase II firm j perceives the reason for firm i's initiating move to P_2 and includes the larger k_i' in its first-order condition (8). Every sale firm j takes from firm i causes firm i to lose less profit than before, so R_j shifts to R_j' - a shift due solely to rivalrous consonance and varying directly with the size of Θ_{ji}. Also, the more sensitive firm i's sales to a reduction in p_j (i.e., the larger b_{ij}) and the more sensitive its own sales

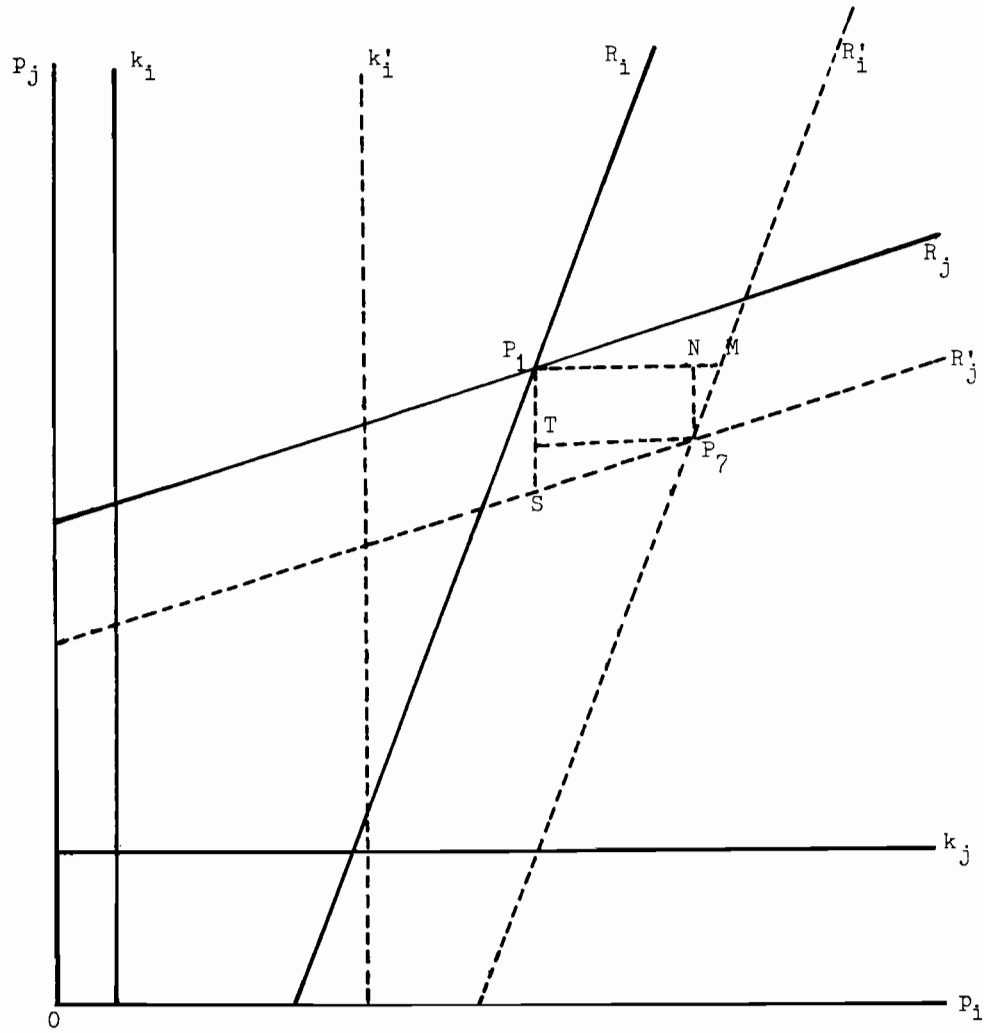


Figure 3 The shift-and-slope analysis of price change

are to p_j, the less will p_j have to be reduced to compensate for the fall in extended profits. Firm j, therefore, moves P from P_4 to P_5, and interactions along the new reaction functions will carry the new equilibrium price to P_7.

Rivalrous consonance then has two opposing price tendencies. Because of the increased *slope* it gives reaction functions, prices tend to increase with rivals' cost increases. But a shift component which is similar to a change in Θ_{ji} caused by a decline in rival profits exerts downward pressure on prices.

On Figure 3 we analyze the net resultant of these two components. For p_i, the change from p_i^1 to p_i^7 is equal to the amount of the shift in R_i (P_1M) less MN, which is simply the slope of R_i' times the change in $p_j(P_1T)$. The same reasoning holds for p_j^1-- p_j^7. That is,

$$\begin{aligned} &1.\ dp_i = \text{shift of } R_i + \text{slope of } R_i \text{ x } dp_j \\ &2.\ dp_j = \text{shift of } R_j + \text{slope of } R_j \text{ x } dp_i \end{aligned} \tag{13}$$

From (12)

$$\begin{aligned} &1.\ dp_i = 0.5dk_i + [0.5(b_{ij} + \Theta_{ii}b_{ji})/b_{ii}]dp_j \\ &2.\ dp_j = (-0.5\Theta_{ji}b_{ji}/b_{jj})dk_i + (0.5(b_{ji} + \Theta_{jj}b_{ij})/b_{jj})dp_i \end{aligned}$$

or

$$\begin{aligned} &1.\ dp_i/dk_i = (2b_{ii}b_{jj} - \Theta_{ji}b_{ij}(b_{ij} + \Theta_{ij}b_{ji}))/D^{\Theta} \\ &2.\ dp_j/dk_i = (b_{ii}(b_{ij} - \Theta_{ji}b_{ij})/D^{\Theta}). \end{aligned} \tag{14}$$

Can rivalvous consonance lead to the perverse case where $dp_i/dk_i < 0$? Because $D^{\Theta} > 0$ by the strict concavity of the objective functions, the sign of (14.1) depends upon the sign of the numerator. It is negative if and only if

$$\Theta_{ji}b_{ij}(b_{ij} + \Theta_{ij}b_{ji}) > 2b_{ii}b_{jj}. \tag{15}$$

From strict concavity the additional necessary condition is that

$$2b_{ii}b_{jj} > 0.5(b_{ij} + \Theta_{ij}b_{ji})(b_{ji} + \Theta_{ji}b_{ij}] \tag{16}$$

Together (15) and (16) imply that in the perverse case,

$$\Theta_{ji}b_{ij} > b_{ji}\,, \tag{17}$$

which in turn implies, because Θ_{ji} is in the unit interval, that $b_{ij} > b_{ji}$.

We will show that the perverse case is possible but unlikely. Define the other-price coefficients as multiples of own-price coefficients:

$$b_{ij} = m_i b_{ii},\ b_{ji} = m_j b_{jj}. \tag{18}$$

Then (15) reduces to

$$b_{ii}/b_{jj} > (2 - \Theta_{ij}\Theta_{ji}m_im_j)/\Theta_{ji}m_i^2\,, \tag{19}$$

and by substituting (18) into (17), the condition of (19) can be tightened to

$$b_{ii}/b_{jj} > 2/m_i^2. \tag{20}$$

Perversity can occur in a duopoly where one rival controls a much larger share of the market than the other; if $m_i < 1$, as we expect for a dominant rival, that necessary predominance may be a larger multiple of 2. Moreover, for the perverse result the weak rival's position is reflected in a very high rival's profit component (i.e., $\theta_{ji}b_{ji}$) which leads to a large shift downward in R_j. Also, the small absolute cross-price sensitivity of the weak rival to the strong rival's price actions insures that the slope of R_i is steep.

The improbability of these conditions in a closed duopoly hinges upon the notion that if a rival with very large sales lost a large quantity through a rise in its price, a sole small rival will fail to benefit from the action by experiencing a large increase in sales. A great leakage from the system is implied. Therefore, the possibility seems most likely when the primary competition of a strong duopolist exists abroad among exogenous exporters and a weak domestic rival acts under strong consonance motivation to exploit a large rise in rival cost.

The improbability of its occurrence for rivals of approximately equal size can be investigated for the case where $b_{ij} = b_{ji} = b$ and $k_i = k_j = k$. Perversity has the best chance of occurring when (from (15)) $\theta_{ij} = \theta_{ji} = 1$, and we adopt that value. Then, from (15) and (16), we obtain the contradiction

$$c > b > c. \tag{21}$$

Hence, the perverse case cannot arise under the best conditions. Even the condition derived from (15) alone implies the realistically unacceptable proposition that cross-price sensitivity is larger than own-price sensitivity.

On the other hand, a rise in k_j may move p_j up or down depending on the numerator of (14.2). Specifically, p_j will fall if

$$b_{ij} < \theta_{ji}b_{ij}. \tag{22}$$

The downward shift of R_j to R_j' in Figure 2 outweighs the upward pull exerted by the slope of R_i'. Note from (17) that if the perverse case of dp_i/dk_i occurs, p_j must also fall, but the converse is not true.

The negative net impact upon prices that rivalrous consonance can inspire is not as paradoxical as might appear at first sight. A fall in the profit margin of a rival can be viewed as a reduction of his power base and will lead a rival to lower the price protection previously afforded. This will be tempered, and in the normal case outweighed, by the rightward shift of firm j's sales function set off by the initial rise in p_i. But where cross-price sensitivities are weak, the net effect of a fall in p_j may occur.

5. CONCLUSION

Rivalrous consonance, with or without constraints, is one path toward a necessary integration of rivalry and cooperation in oligopoly theory. The theory isolates an interesting Nash equilibrium at the generalized Stackelberg point. And it casts light on the contradictory impulses that a unilateral cost increase inspires among oligopolists. On the one hand the unfortunate rival's attempt to pass through the increase by a price rise leads rivals to increase their own prices as demand strengthens.But the decline in the initiating

firm's profit margin leads rivals to reduce the price protection formerly afforded it. Net price changes in the industry are the net resultants of these two forces.

REFERENCE

Kuenne, R.E., 1986, *Rivalrous Consonance: A Theory of General Oligopolistic Equilibrium*, North-Holland, Amsterdam.

PART B

SPATIAL INTERACTION

Advances in Spatial Theory and Dynamics
A.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 9

On the Spatiotemporal Dynamics of Capital and Labour

T. Puu

1. INTRODUCTION

Hotelling (1921) in an early work formulated a model for population growth and dispersal, assuming a logistic growth combined with linear diffusion from locations of higher to locations of lower population density. This interesting work left no trace in economics, where researchers seem to prefer linear to logistic growth processes, and where the interest in phenomena in the geographical plane has constantly decreased. The same model was reinvented by Fischer (1937) and elaborated by Skellam (1951) in order to deal with nonhuman populations, a context in which it became a great success. Hotelling's work was republished as late as 1978, but obviously it still holds little appeal for the economics profession.

The logistic growth function assumes that there is a given saturation population that nature can support. For a human population, however, the total means of subsistence are not merely provided by nature: to a very large extent they are produced by the population itself. More people in general produce more goods and only eventually do diminishing returns force per capita output down. Moreover, man himself, unlike other animals, decides the per capita living standard at which the population stays stationary. Finally, all these conditions are subject to change over time due to technological developments and changing attitudes.

In a recent publication, Puu (1985), the author tried to cope with these problems by introducing an explicit production function with increasing-decreasing returns. In line with the modern theory of structural change (see Gilmore,1981; or Poston and Stewart, 1978), where truncated Taylor series universally represent qualitative phenomena, the production function was formulated as the lowest-order polynomial that could represent a technology with the characteristics mentioned. It contained two arbitrary constants, representing technological efficiency and optimal scale of operations. Together with the stationary population living standard as a third parameter, they could change over time and thus trigger off sudden growth processes.

Similarly, the spatial movement of population was related to differences in living standards rather than to differences in population density per se. It was then possible to study the character of spatially non-homogeneous (agglomerative) steady states, and it was shown that they became a possibility once technological efficiency had passed a certain threshold.

All the changes in the relation of output to labour input were, however, assumed to be exogenous to the system. The next step of development is natural: to include the accumulation of capital as an explicit process in the model, thus explaining one of the forces for change in the productivity of labour.

Accordingly, a two-input production function is needed. Again we will formulate it as a polynomial of a sufficiently high degree to account for a transition from increasing to decreasing returns. The inputs, labour and capital, have to enter the production function in such a way as to produce a positive synergetic effect, and this can be achieved by a third degree polynomial. As only the qualitative phenomena are of interest it is legitimate to work with such a truncated Taylor series of low order.

The formation process for capital is modelled as follows. The single output is assumed malleable so that it can either be consumed or invested as capital. Expectations of capital income are assumed to be determined, like labour incomes, by marginal productivity and are reinvested. Net additions to capital stock are obtained after deduction of depreciation, assumed proportionate to capital stock. This leads to steady state capital stock when marginal productivity equals the rate of depreciation.

For labour, a supply process of higher order is assumed. The reason for this is that the model must be able to account for population growth at early stages of development when mankind is supported by the yields of Nature without any additional output from organized production.

2. DYNAMICS

Write the production function as

$$q = \alpha(\beta(k+\ell)^2 - k^3 - \ell^3) \tag{1}$$

where k denotes capital, ℓ denotes labour, and q denotes output. This is the simplest possible way of formulating a production function with increasing-decreasing returns to scale as a truncated power series. We note that the term $2k\ell$ accounts for a synergetic effect. In order that decreasing returns should eventually dominate the negative terms must be of the third order.

We note that output has a unique maximum when both marginal productivities are zero. This happens for $k = \ell = 4/3\ \beta$. The transition from increasing to decreasing returns to scale occurs when k and ℓ satisfy the implicit equation $(k^3 + \ell^3)/(k + \ell)^2 = \beta/2$. When the right hand side constant is doubled the same implicit equation represents the points for which output becomes zero. The general features of this production function are shown in Figure 1.

The output shares of capital and labour evaluated at the current marginal productivities are

$$\frac{\partial q}{\partial k}k = 2\alpha\beta(k+\ell)k - 3\alpha k^3 \tag{2}$$

$$\frac{\partial q}{\partial \ell}\ell = 2\alpha\beta(k+\ell)\ell - 3\alpha\ell^3 \tag{3}$$

If the production function were linearly homogeneous these input shares would exhaust output, but given (1) they sum up to more or less, depending on whether production is in the region of increasing or decreasing returns. This makes it problematic to take (2)-(3) as measures of factor incomes. If capital accumulation or population increase is based on factor income the model yields different results depending on which input is assumed to

be remunerated according to the marginal conditions and which takes the residual. Another possibility is to assume, as is done in most of the economics of growth, that a given proportion of output is saved (and invested) and the remaining fraction consumed.

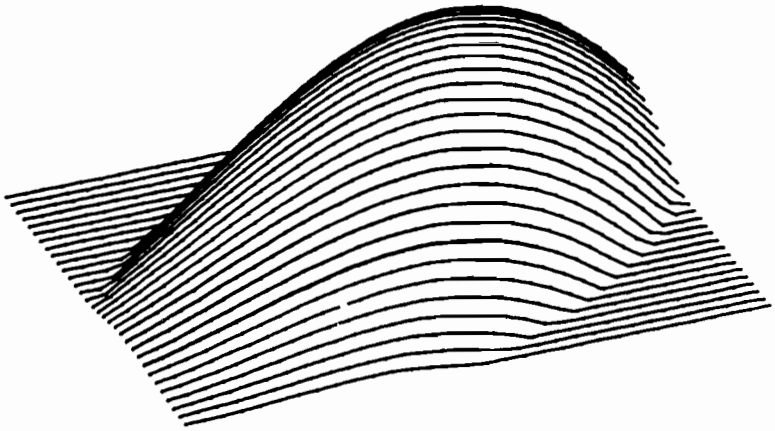


Figure 1 The production function

We shall evade these difficulties by assuming that both capital and labour do estimate their incomes by (2) and (3) respectively. As the estimates are used for formulating a dynamic process involving marginal change it does not matter that the expectations are inconsistent; indeed, it may even seem reasonable to base the estimates on marginal productivities.

Accordingly, the growth of capital stock is assumed to be

$$\dot{k} = 2\alpha\beta(k+\ell)k - 3\alpha k^3 - \kappa k \tag{4}$$

where κ denotes the rate of capital depreciation. It is easy to interpret (4). Estimated capital income is invested and results in an increase of capital stock to the extent that gross investment exceeds depreciation. Capital stock is in equilibrium when marginal productivity equals the rate of depreciation; i.e., when the "golden rule" is satisfied.

For the dynamics of the labour force we assume a growth rate equal to its estimated income share, plus the natural supply of means of subsistence, normalized to unity, minus the subjective "need" for goods $\gamma\ell$.

$$\dot{\ell} = \ell(1+ 2\alpha\beta(k+\ell)\ell - 3\alpha\ell^3 - \gamma\ell) \tag{5}$$

Our first objective is to investigate the dynamics of (4)-(5) before space and diffusion are introduced. Obviously, (4) is zero when $k = 0$, and (5) is zero when $\ell = 0$. But there are more interesting possibilities. In addition (4) is zero whenever k and ℓ satisfy the relation $\ell = (\kappa/2\alpha\beta) - k + (3/2\beta)k^2$, whereas (5) is zero when $k = (\gamma/2\alpha\beta) - (1/2\alpha\beta)/\ell - \ell + (3/2\beta)\ell^2$. The curves for zero growth of capital and labour are easily sketched as in each case one of the variables can be solved as an explicit function of the other. The zero curve for capital is simply a quadratic with a minimum, whereas that for labour has a more complex character, possibly having a point of inflection and a backward-bending section. As indicated in Figure 2, the pair of zero curves may have up to four intersections. These, of course, are the critical (or equilibrium) points of the system. The corresponding flow portrait is shown in Figure 3. Two of the intersection points are stable nodes and two are saddles. Together with the stable node on the vertical axis this makes up three different zones of attraction in the phase space.

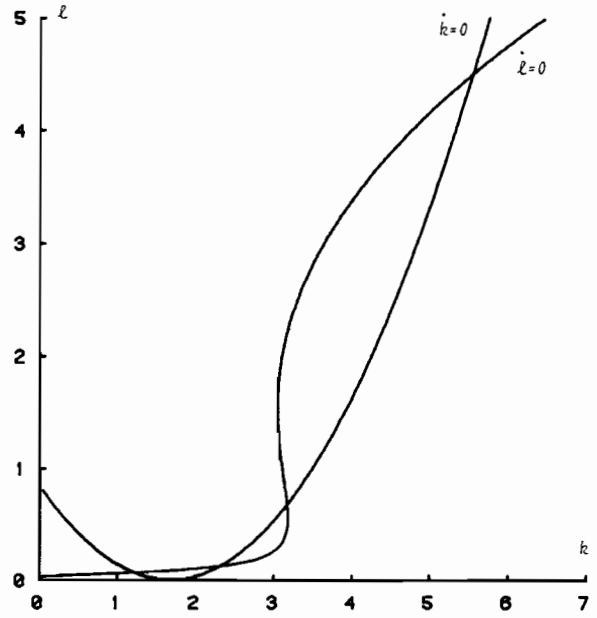


Figure 2 Zero lines for capital and labour growth

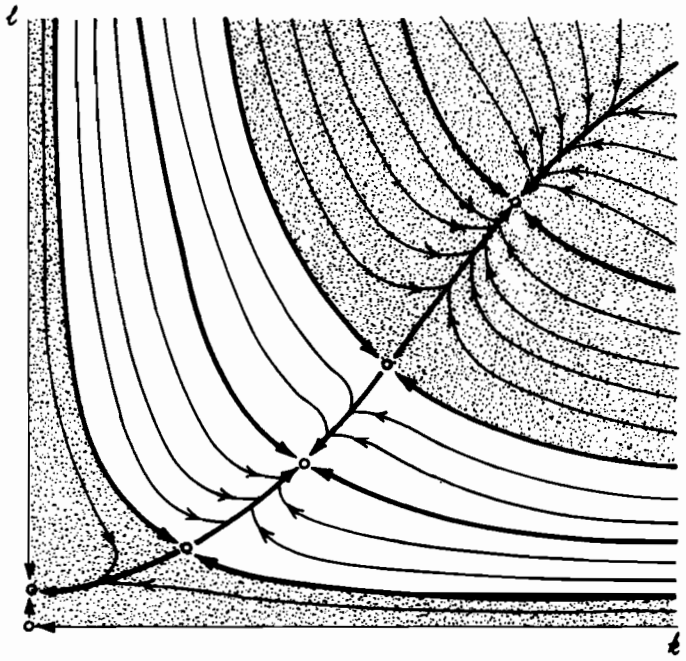


Figure 3 Flow portrait with three attraction zones

By various parameter changes - for instance, by the variation of the depreciation constant κ - a saddle point may fuse with a neighbouring node. If this occurs, the latter disappears as a possibility for stable equilibrium, and its former zone of attraction fuses with the one on the side of the saddle point. In Figure 4 three different fusions are illustrated for three different positions of the parabola, and the corresponding phase portraits are displayed in Figures 5 a-c. By studying these diagrams it is easy to understand how new equilibria emerge and old ones are destroyed as a result of smooth parameter changes, which thus may trigger off sudden growth or decline in population and capital stock in an otherwise smooth development.

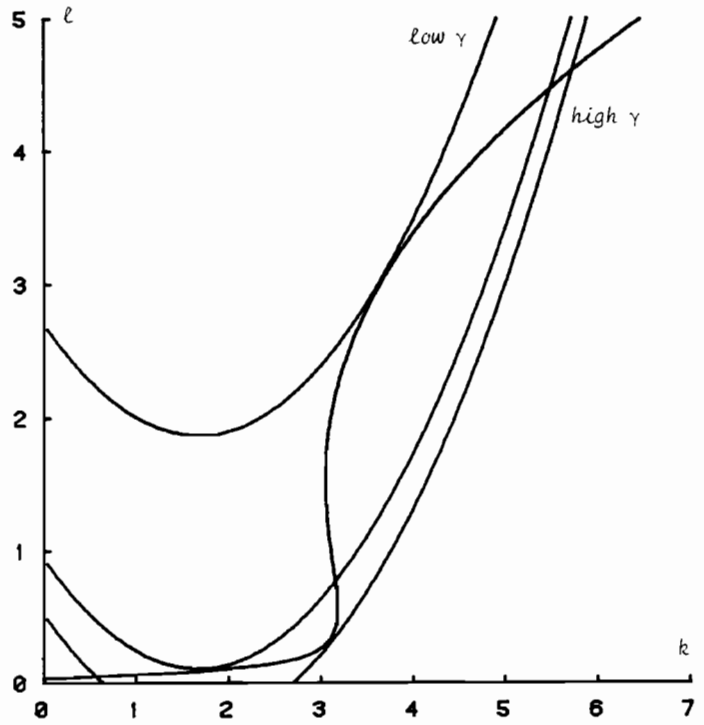


Figure 4 Zero lines in degenerate cases

3. COMPUTATIONAL ASPECTS

Analytically, we may study the behaviour in the neighbourhood of an equilibrium by linearization. Linearization of (4)-(5) in the neighbourhood of some equilibrium with nonzero $\bar{k}$, $\bar{\ell}$, produced the system:

$$\dot{k} = (2\alpha\beta\bar{k} - 6\alpha\bar{k}^2)(k-\bar{k}) + 2\alpha\beta\bar{k}(\ell-\bar{\ell}) \tag{6}$$

$$\dot{\ell} = 2\alpha\beta\bar{\ell}^2(k-\bar{k}) + (2\alpha\beta\bar{\ell}^2 - 6\alpha\bar{\ell}^3 - 1)(\ell-\bar{\ell}) \tag{7}$$

where we have used (4)-(5) to eliminate the constants κ and γ.

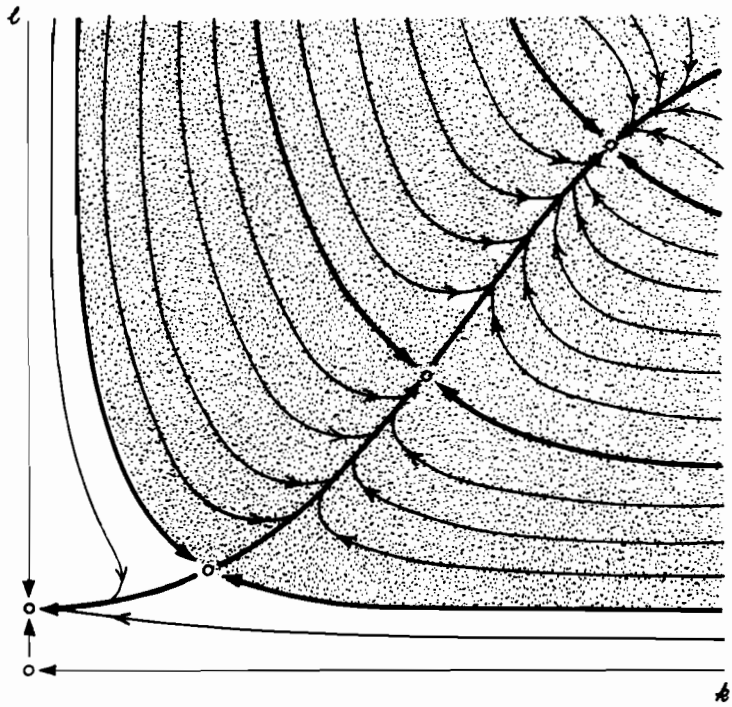


Figure 5a Degenerate flow portrait with extended upper attraction region

Denoting the eigenvalues of this system by μ_1 and μ_2 we have for the trace and determinant

$$\text{Tr} = \mu_1 + \mu_2 = 2\alpha((\beta-3\bar{k})\bar{k} + (\beta-3\bar{\ell})\bar{\ell}^2 - 1 \tag{8}$$

$$\text{Det} = \mu_1\mu_2 = -2\alpha((\ell+6\alpha\bar{\ell}^3)(\beta-3\bar{k}) + 6\alpha\beta\bar{k}\,\bar{\ell}^2)\bar{k} \tag{9}$$

The discriminant can then be computed:

$$\Delta = (\text{Tr})^2 - 4\,\text{Det} = (2\alpha\beta\bar{k} - 6\alpha\,\bar{k}^2 - 2\alpha\beta\bar{\ell}^2+6\alpha\bar{\ell}^3+1)^2 + 16\alpha^2\beta^2\bar{k}\,\bar{\ell}^2 > 0. \tag{10}$$

From (10) we see that the eigenvalues are always real, i.e., nodes and saddles if they are not degenerate. Degeneracy implies a zero determinant. We see that the determinant is always negative if $k < \beta/3$. In that case the eigenvalues have opposite signs and the equilibrium is a saddle point. For $k > \beta/3$ the determinant can become positive, so that the critical point is a node. In that case there is a tendency for the trace to be negative, and this explains the appearance of stable nodes along with the saddles in the illustrations.

To illustrate the computational aspects suppose we have $\alpha=1/6$, $\beta=2$. Then (4)-(5) obviously has an equilibrium point in $\bar{k}=\bar{\ell}=1$. Implicitly the remaining constants are $\kappa=5/6$, and $\gamma=11/6$. If we insert these numbers in the linearized equations (6)-(7) we get the coefficient matrix

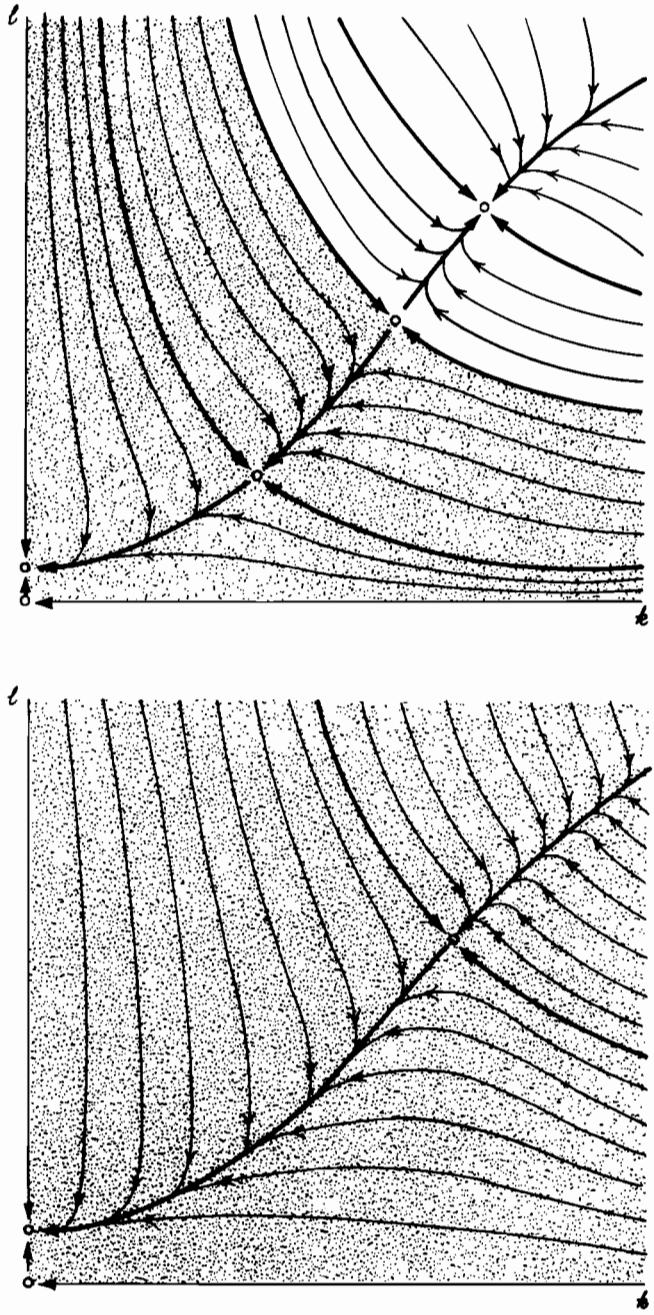


Figure 5 b-c Degenerate flows with extended lower attraction regions

$$\begin{bmatrix} -\frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{4}{3} \end{bmatrix} = A$$

Its determinant is zero, so the critical point is degenerate. The eigenvalues are -5/3 and 0 respectively, and the corresponding right hand eigenvectors can easily be computed as $(-1/\sqrt{5}, 2/\sqrt{5})$ and $(2/\sqrt{5}, 1/\sqrt{5})$. Also, the left hand eigenvectors turn out to be identical with the right hand ones. Accordingly the matrix $(A - \mu I)$ happens to be self-adjoint in our example.

The linearized equations (6)-(7) with the numerical matrix elements from our example inserted form a linear system with constant coefficients, whose solution

$$k = 1 - a \exp(-5t/3)/\sqrt{5} + 2b \exp(0t)/\sqrt{5} \tag{11}$$

$$\ell = 1 + 2a \exp(-5t/3)/\sqrt{5} + b \exp(0t)/\sqrt{5} \tag{12}$$

is trivial from the elementary theory of differential equations. Instead of paying further attention to this we define the new variables $\xi(t) = a \exp(-5t/3)$ and $\eta(t) = b \exp(0t) = b$. Using the method suggested by Haken (1977, 1983) we set

$$k = 1 - \xi/\sqrt{5} + 2\eta/\sqrt{5} \tag{13}$$

$$\ell = 1 + 2\xi/\sqrt{5} + \eta/\sqrt{5} \tag{14}$$

where we now regard the amplitudes a and b as time-dependent. Thus (13)-(14) is a linear change of coordinates from k, ℓ to ξ, η, where capital and labour start out from their equilibrium position (1,1) in the characteristic directions as ξ or η start to change from zero. The expressions (13)-(14) are now inserted into the original nonlinear pair of differential equations (4)-(5), where the numerical values of the constants from the example are substituted. These equations then express $\dot{k}$ and $\dot{\ell}$ as functions of the new variables ξ, η involving terms up to cubic or quartic respectively. It should be observed that these new differential equations are exact, since they only retain information concerning the characteristic directions from the linear approximation.

Now, differentiating (13)-(14) we get

$$\dot{k} = -\dot{\xi}/\sqrt{5} + 2\dot{\eta}/\sqrt{5} \tag{15}$$

$$\dot{\ell} = 2\dot{\xi}/\sqrt{5} + \dot{\eta}/\sqrt{5} \tag{16}$$

which can be diagonalized by using the left hand eigenvectors. Making use of the derived expressions for $\dot{k}$ and $\dot{\ell}$ we finally obtain

$$\dot{\xi} = -\frac{5}{3}\xi - \frac{123\xi^2+86\xi\eta-9\eta^2}{30\sqrt{5}} - \frac{157\xi^3+132\xi^2\eta+16\xi\eta^2-24\eta^3}{150} - \frac{96\xi^4-192\xi^3\eta+144\xi^2\eta^2+48\xi\eta^3+6\eta^4}{150\sqrt{5}} \tag{17}$$

$$\dot{\eta} = -\frac{94\xi^2-32\xi\eta+9\eta^2}{30\sqrt{5}} - \frac{74\xi^3+116\xi^2\eta-52\xi\eta^2-52\eta^3}{150}$$
$$-\frac{48\xi^4+96\xi^3\eta^2+72\xi^2\eta^2+24\xi\eta^3+3\eta^4}{150\sqrt{5}} \tag{18}$$

Observe once again that these differential equations are exact. They follow from the nonlinear system using the transformation of coordinates. The new equations look rather awkward, but they have the advantage that they are diagonalized in the linear terms. Moreover, the coefficients of the linear terms are the eigenvalues, -5/3 and 0.

As there is linear damping in (17) we may assume that the system is stable in the variable ξ, i.e., we can equate (17) to zero, thus obtaining an implicit equation relating ξ and η. This relation can be interpreted to provide the value of the former as dependent on the latter, so that ξ might be eliminated from (18) which turns into a differential equation for the slowly changing mode η alone. We shall not follow this use of the "slaving principle" or "adiabatic elimination" any further. Details can be found in Haken (1977, 1983).

The value of the example is that it illuminates the manner in which computations are carried out in systems like this. Because of the nonlinearities the arithmetic labour can become formidable. The value of the slaving principle lies in the fact that in the study of instabilities the stable modes can be neglected and treated as instantaneously damped. In order to find out which modes are stable and which are unstable the Haken method of diagonalization of nonlinear systems is needed, because the original variables seldom fall into stable or unstable categories as they appear.

4. DIFFUSION

Let us now introduce space and diffusion. Suppose there is a tendency for the marginal productivity of capital and labour to be equalized over space. If there are spatial differences in marginal factor productivities, factors are assumed to flow from locations of low to locations of high marginal productivity.

Using the commonplace symbol $\nabla^2 = (\partial^2/\partial x^2+\partial^2/\partial y^2)$ for the Laplacian measure of spatial differences, x and y denoting the space coordinates, we introduce the diffusion terms

$$-\delta\nabla^2(2\alpha\beta(k+\ell) - 3\alpha k^2) \tag{19}$$

$$-\delta\nabla^2(2\alpha\beta(k+\ell) - 3\alpha\ell^2) \tag{20}$$

in the right hand sides of equations (4) and (5) respectively. Next, assume that the functions to which the Laplacian operator is applied solve the eigenvalue problem

$$\nabla^2 S_i(x,y) = -\lambda_i S_i(x,y) \tag{21}$$

where the $S_i(x,y)$ are the eigenfunctions associated with the shape of the region in space. When the marginal productivities of capital and labour vary over space according to any of the eigenfunctions, then (19)-(20) can be replaced by

$$\delta\lambda_i(2\alpha\beta(k+\ell) - 3\alpha k^2) \tag{22}$$

$$\delta\lambda_j(2\alpha\beta(k+\ell) - 3\alpha\ell^2) \tag{23}$$

where the spatial operators are no longer present. We have chosen different indices for the two eigenvalues in order to stress that the marginal productivities of capital and labour need not vary in accordance with the same eigenfunction.

We now want to study the linear stability of a spatially homogeneous stationary solution. In the case of spatial homogeneity the diffusion terms vanish and we are left with the original system (4)-(5). Accordingly, the same stationary solutions $\bar{k}$ and $\bar{\ell}$ apply as before. In order to investigate stability in the neighbourhood of this solution we introduce the linearization (6)-(7), but now we have to add linearizations of the diffusion terms (22)-(23). This transforms (6)-(7) into

$$\dot{k} = (\bar{k}+\delta\lambda_i)(2\alpha\beta-6\alpha\bar{k})(k-\bar{k}) + (\bar{k} + \delta\lambda_i)2\alpha\beta(\ell-\bar{\ell}) \tag{24}$$

$$\dot{\ell} = (\bar{\ell}^2+ \delta\lambda_j)2\alpha\beta(k-\bar{k}) + ((\bar{\ell}^2+ \delta\lambda_j)(2\alpha\beta-6\alpha\bar{\ell}) -1)(\ell-\bar{\ell}). \tag{25}$$

Let us find out how the determinant and trace of this new system differ from those of the system (6)-(7). In order to distinguish the new ones we denote them Tr* and Det*. Some computation shows that

$$Tr^* = Tr + \delta\lambda_i\, 2\alpha(\beta-3\bar{k}) + \delta\lambda_j\, 2\alpha(\beta-3\bar{\ell}) \tag{26}$$

and

$$Det^* = (1+ \delta\lambda_i/\bar{k})((\delta\lambda_j/\bar{\ell}^2)2\alpha\bar{k}(\beta-3\bar{k}) + Det). \tag{27}$$

The diffusion constant δ is positive as before, and so are the eigenvalues λ_i and λ_j as can be demonstrated by applying Green's first identity to (21) multiplied by S_i. See Courant and Hilbert (1953). Finally, the equilibrium values $\bar{k}$ and $\bar{\ell}$ are also positive. We see that the determinant is increased or decreased with ascending eigenvalues λ_j depending on whether $\bar{k}$ is smaller or larger than the critical value $\beta/3$.

Thus, if capital stock in a spatially homogeneous stationary equilibrium is larger than a certain critical value, then the determinant of (24)-(25) decreases for increasingly nonhomogeneous modes of spatial variation. Put succinctly, the spatially homogeneous solution is unstable if capital stock exceeds a certain critical value. This is in line with the conclusion in Puu (1985) that exceeding a threshold of technological efficiency conditions agglomerative patterns. Obviously the same role is played by a sufficiently large accumulation of capital.

REFERENCES

Courant, R. and D. Hilbert, 1937; 1953, *Methods of Mathematical Physics*. Interscience, New York. (Translated from the German original of 1937).

Fisher, R.A., 1937, "The Wave of Advance of Advantageous Genes", *Annals of Eugenics*, 7:355-369.

Gilmore, R., 1981, *Catastrophe Theory for Scientists and Engineers*, Interscience, New York.

Haken, H., 1977, *Synergetics, An Introduction*, Springer Verlag, Berlin.

Haken, H., 1983, *Advanced Synergetics, Instability Hierarchies of Self-Organizing Systems and Devices*, Springer Verlag, Berlin.

Hotelling, H., 1921; 1978, "A Mathematical theory of Migration," M.A. Thesis presented at the University of Washington, republished in *Environment and Planning A*, 10:1223-1239.

Poston, T. and I. Stewart, 1978, *Catastrophe Theory and its Applications*, Pitman, London.

Puu, T., 1985, "A Simplified Model of Spatiotemporal Population Dynamics", *Environment and Planning A*, 17:1263-1269.

Skellam, J.G., 1951, "Random Dispersal in Theoretical Populations", *Biometrika*, 38:196-218.

Advances in Spatial Theory and Dynamics
A.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 10

Finite Spectral Analysis of Multiregional Time Series

T.E. Smith

1. INTRODUCTION

Applications of spectral analysis to multiregional time series now abound in the literature, and have been well summarized by Bennett (1979) and others. But unlike more traditional multivariate techniques, such as regression and correlation analysis, the techniques employed in multivariate spectral analysis are relatively sophisticated in nature. Moreover, while many systematic developments of this theory exist in the literature (including Jenkins and Watts (1968), Hannan (1970), Dhrymes (1970), Brillinger (1975), and Fuller (1976), among others), the underlying mathematical theory of stochastic (Fourier) integrals is unfamiliar to most potential users. In fact, the fundamental *spectral representation theorem* for stationary processes, which forms the corner stone of the entire theory, can only be proved by methods which are beyond the scope of most standard texts on the subject (see for example the proofs in Hannan (1970; Theorem 2.3.2") and in Cramer and Leadbetter (1967; sections 7.5 and 8.1)). With this in mind, the central objective of this paper is to develop a conceptually simpler and more heuristic approach to multivariate spectral analysis.

This approach, which we designate as *finite spectral analysis*, essentially reduces the classical infinite-dimensional formulation of spectral analysis to a finite-dimensional version of principal components analysis. Within this finite-dimensional framework, the spectral representation theorem together with its analytical consequences can be derived in an elementary manner (requiring only matrix algebra). The mathematical possibility of such a reduction is by no means new to the literature. In particular, it has long been known that the principle components generated by finite segments of (absolutely summable) stationary processes converge to the spectral representation of the entire process as the segment becomes arbitrarily large (see for example Brillinger (1975; section 4.7) and Fuller (1976; sections 4.2 and 7.4)). The standard proof of this result employs an approximation of stationary covariance matrices by 'circulant' matrices which can be spectrally decomposed in a simple way. Moreover, it is also well known that such matrices correspond to the covariance structures of stationary periodic processes (see for example Jenkins and Watts (1968; section 11.1.2), Anderson (1971; section 6.5.2), and Streitberg (1979; section 6.2.3.3)). These stochastic processes, which we designate as *finitely-stationary processes*, are of central importance in the present approach. In particular, it will be argued that such processes constitute the simplest stationary processes which can capture all the statistical covariation within a given finite stationary sample sequence, or *stationary segment*. Moreover, since all actual observations on

stationary processes are restricted to such segments, it is appropriate to seek the simplest model of the unknown process which is consistent with this observed information. Thus, our approach is to model an infinite stationary process in terms of the finitely-stationary process which 'best' represents its observable stationary segment.

Given this objective, our central theoretical result is to show that there exists a unique (normalized) finitely-stationary process which best represents a given stationary segment in terms of the 'max-min correlation' between their corresponding components. This process, designated here as the *circular smoothing* of a stationary segment, is essentially a periodic moving average of the underlying segment. The spectral representation of this process is readily derivable in closed form, and is shown to yield a simple geometric model for studying the possible cyclical behavior implied by the observed segment. Moreover, the statistical 'spectrum' of this process is shown to correspond precisely to the *periodogram* of the original segment. Since the sample form of this periodogram constitutes the basic set of statistics upon which all standard spectral estimators are based, these results yield a new and direct theoretical interpretation of classical spectral analysis. Hence, this finitely-stationary model of infinite processes provides all the theory needed for actual calculations - in a simple and easily intelligible form. In addition, these results provide a new theoretical underpinning for the many practical applications of finite Fourier transforms to the approximation of stationary covariance structures, as in the methods of 'spectral regression' studied by Duncan and Jones (1966), Engle (1974), and Harvey (1978).

In order to motivate the basic rationale for this finite approach, it is convenient to begin in the next section with an informal development and discussion of the main results. The formal development will begin in Section 3 with a general analysis of finitely-stationary processes, including both the real and complex spectral representations of such processes. These representations are interpreted in geometric terms, and are compared to the ordinary principle components for such processes. The main results of the paper are developed in Section 4, where circular-smoothing processes are developed formally and shown to be optimal representatives of stationary segments. The results of Section 3 are then applied to these processes to show that their spectra correspond to the periodograms of the segments which they represent.

2. MOTIVATION AND DISCUSSION OF RESULTS

Consider a sequence of random variables Y_{tj} describing the behavior of a given regional attribute (employment level, per capita income level, etc.) within a system of regions j=1,..,n over time periods $t \in \mathbf{T} = \{...,-1,0,1,...\}$. If $Y_t' = (Y_{t1},..,Y_{tj},..,Y_{tn})$ denotes the associated random vector of regional variates at time t,[1] then the resulting sequence of random vectors $\{Y_t \mid t \in \mathbf{T}\}$ over time defines an n-dimensional stochastic process, or *n-process*. Next, letting $m(t) = E(Y_t)$ denote the *mean* of Y_t and $X_t = Y_t - m(t)$ denote the *residual* of Y_t about its mean, we may express Y_t as the sum of two components:

$$Y_t = m(t) + X_t \quad , \quad t \in \mathbf{T} \tag{2.1}$$

Hence, the n-process $\{Y_t \mid t \in \mathbf{T}\}$ is seen to be decomposable into a deterministic *trend sequence* $\{m(t) \mid t \in \mathbf{T}\}$ and a stochastic *residual process* $\mathbf{X} = \{X_t \mid t \in \mathbf{T}\}$. In terms of this decomposition our present interest focuses exclusively on the stochastic structure of the residual process **X**.

---

[1] Since the present analysis is meaningful for any random vector of attributes, one may easily consider multiple regional attributes by letting each Y_{tj} denote a vector of regional attributes Y_{tjk}. However, for notational simplicity, we shall treat each regional component as a single random variable.

To analyze this process directly, it is convenient to assume that the trend term m(t) is either known or can be determined explicitly, and hence that the residuals $X_t = Y_t - m(t)$ can be treated as observable quantities. As a next step in the analysis of **X**, observe that by definition the sequence of residual variates $\{X_t \mid t \in \mathbf{T}\}$ is *zero - mean stationary* in the sense that $E(X_t) = E(Y_t) - m(t) = 0$ for all $t \in \mathbf{T}$. Our major statistical hypothesis about **X** is that its covariance structure also remains stationary over time. More precisely, it is hypothesized that the covariance between any residuals X_t and $X_{t+\tau}$ depends only on the time lag τ between them, i.e. that for all t, v, t', v' $\in \mathbf{T}$ with $t-v = t'-v' = \tau$.

$$\text{cov}(X_t, X_v) = \text{cov}(X_{t'}, X_{v'}) = K(\tau) \,. \tag{2.2}$$

Under this hypothesis, **X** is said to be *stationary* (i.e. covariance stationary and zero-mean stationary) with *covariance kernel* $K = \{K(\tau) \mid \tau \in \mathbf{T}\}$.

2.1 The Finite-Sample Problem

Even assuming that each residual variate X_t is potentially observable, it is clear that the process **X** can never be observed in its entirety. In particular, for each sample sequence $X_t,...,X_{t+T}$ which can actually be observed, the unknown process **X** can only be viewed through the finite *sample window* [t,..,t+T] shown below:

$$\mathbf{X} = \{..., X_{t-1}, \underbrace{[X_t,..,X_{t+T}]}_{\text{sample window}}, X_{t+T+1},...\}$$
$$(?) \qquad\qquad (?)$$

Thus, from a practical viewpoint, the most basic problem to be addressed is how to treat the rest of the process **X** *outside* this sample window. Our approach to this problem marks a fundamental point of departure from more standard methods of analyzing stationary processes. The essence of this approach is to discard the unobservable portion of **X** and to construct a simpler stationary model of **X** which involves only the information contained in the observable portion of **X**.

2.2 Finitely-Stationary Processes

To motivate this approach, we begin by considering the possible covariance information which is contained in a given finite segment of **X**. If $\mathbf{X}_T = [X_0,..,X_T]$ denotes a given *stationary segment* from **X**, then observing that $\text{cov}(X_\tau,X_0) = K(\tau)$ and $\text{cov}(X_0,X_\tau) = K(-\tau)$ for each $\tau = 0,1,..,T$, we may conclude that $\mathbf{X}_T$ yields information about the finite segment [K(-T),..,K(0),..,K(T)] of the covariance kernel K for **X**. This in turn implies from the definition of stationarity that for *any* time period $t \in \mathbf{T}$, the segment $\mathbf{X}_T$ yields information about the covariance relationships between X_t and those variates within T time periods of t, i.e. within a *covariance window* [t-T,..,t,..t+T] about t, as shown below:

$$\mathbf{X} = \{..., X_{t-T-1}, \underbrace{[X_{t-T},..,X_t,..,X_{t+T}]}_{\text{covariance window}}, X_{t+T+1},...\}$$
$$(?) \qquad\qquad (?)$$

On the other hand, the joint distribution of the random vectors in $\mathbf{X}_T$ can never yield information about the statistical relations between X_t and any variates outside this

covariance window. Thus, to capture *all* the covariance information contained in $\mathbf{X}_T$, it suffices to consider a representative covariance window in $\mathbf{X}$.

With this in mind, we now consider stationary models of $\mathbf{X}$ in which all relevant stochastic behavior can be represented by an arbitrary covariance window, i.e. by an arbitrary segment of 2T + 1 time periods. To do so, observe first that the only stochastic processes $\mathbf{Z} = \{Z_t \mid t \in \mathbf{T}\}$ which can possibly be represented by a finite segment $[Z_{t-T},..,Z_t,..,Z_{t+T}]$ are those 'periodic' processes which amount simply to repetitions of this segment, i.e. in which all random variates 2T + 1 time periods apart are identical. To make this notion precise, it is convenient to introduce the following equivalence relation over time periods. If we now denote the particular interval of 2T + 1 time periods centered about the origin by [T] = [-T,..,0,..,T] then for each time period $t \in \mathbf{T}$ we may define the T-*equivalent* $\langle t \rangle$ of t to be the unique element of [T] which differs from t by an integral multiple of 2T + 1.

In these terms, an n-process $\mathbf{Z} = \{Z_t \mid t \in \mathbf{T}\}$ is said to be T-*periodic* if and only if $X_t = X_{\langle t \rangle}$ for all $t \in \mathbf{T}$. Hence each T-periodic process $\mathbf{Z}$ is completely representable by any segment of 2T + 1 time periods, and in particular, by the centered segment $\mathbf{Z}[T] = [Z_{-T},..,Z_0,..,Z_T]$. With this definition, it follows that the models we seek for the unknown stationary process $\mathbf{X}$ consist precisely of those stationary processes which are also T-periodic. We now designate such processes as T-*stationary processes*,[2] and denote the class of such processes by $\mathbf{P}_T$. When the specification of T is not relevant, we speak simply of *finitely-stationary processes*.

To clarify this notion of finite-stationarity further, it is instructive to observe that even if a T-periodic process $\mathbf{Z}$ is generated by a stationary segment $\mathbf{Z}[T] = [Z_{-T},..,Z_0,..Z_T]$, the process $\mathbf{Z}$ itself need *not* to be stationary. For if $\mathbf{Z}[T]$ is a segment from a stationary process with covariance kernel K, then observing that $\langle T+1 \rangle = -T$, we set that $\mathbf{Z}$ can only be stationary if $K(2T) = \text{cov}(Z_T,Z_{-T}) = \text{cov}(Z_T,Z_{T+1}) = \text{cov}(Z_0,Z_1) = K(-1)$. But since this identity fails to hold for arbitrary covariance kernels K, we may conclude that something more is implied by finite-stationarity. The additional conditions which will ensure T-stationarity of $\mathbf{Z}$ may be clarified in geometric terms by wrapping the interval [T] around the circle, as shown in Figure 2.1.

This figure shows that adjacencies among the random vectors in a T-periodic process $\mathbf{Z}$ correspond precisely to adjacencies on the circle. For example, the random vectors Z_T and Z_{T+2} in $\mathbf{Z}$ are two time periods apart, and their corresponding positions T and $1\text{-}T = \langle T+2 \rangle$ are two positions apart on the circle in Figure 2.1. Hence, we see that T-stationarity of $\mathbf{Z}$ corresponds to 'circular' stationarity of its finite representation $\mathbf{Z}[T] = [Z_{-T},..,Z_0,..Z_T]$. In view of this circularity property, finitely-stationary processes are often referred to as 'circular processes' in the literature (see for example Anderson (1971) and Streitberg (1979)).

2.3 Circular-Smoothing Processes

From among all possible T-stationary processes $\mathbf{Z} \in \mathbf{P}_T$ which could be employed to model $\mathbf{X}$, we seek to identify a unique process which in some sense 'best' represents the observable portion $\mathbf{X}_T$ of $\mathbf{X}$. To do so, let us first recall that since the desired T-stationary process $\mathbf{Z}$ is intended to model an arbitrary covariance window [t-T,..,t.,,.t+T], the actual position of the time interval [0,..,T] in $\mathbf{Z}$ is completely arbitrary. This means that in order to determine how well a given process $\mathbf{Z}$ represents $\mathbf{X}_T = [X_0,..,X_T]$, we must take into account all possible positions which $\mathbf{X}_T$ could occupy in $\mathbf{Z}$. As can be seen from the

[2]Note that since the actual period of repetition in a T-periodic process is the full length of the interval [T], it would be more precise to refer to [T]-periodic and [T]-stationary processes. However, for notational simplicity, we choose to drop the brackets.

circular representation of **Z** in Figure 2.1, there are only 2T + 1 distinct positions which the sequence $[X_0,..,X_T]$ could occupy in **Z**, namely one for each possible location of X_0

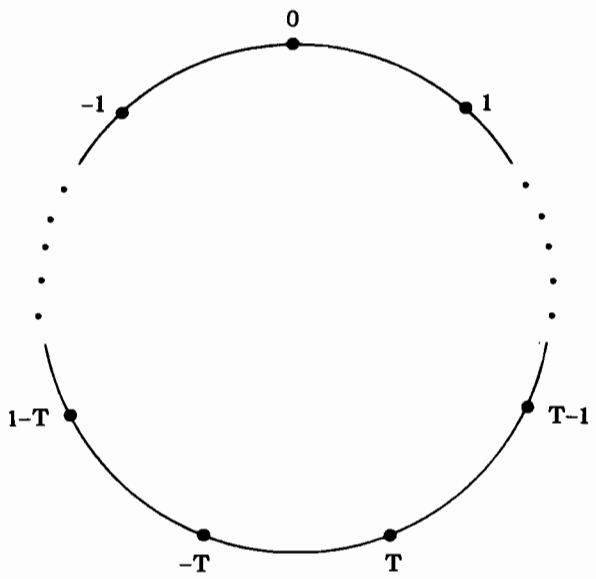


Figure 2.1 Circular Adjacencies

on the circle.[3] In order to compare **Z** with each of these positional possibilities for $\mathbf{X}_T$, it is convenient to construct a set of independent *positional replicates* $\mathbf{X}_T^\tau = [X_0^\tau,..,X_T^\tau]$ of $\mathbf{X}_T$, one for each position as shown in Figure 2.2. By correlating each replicate $[X_0^\tau,..,X_T^\tau]$ with its corresponding segment $[Z_{\langle\tau\rangle},..,Z_{\langle\tau+T\rangle}]$ in **Z**, we will then be able to measure how well **Z** agrees with $\mathbf{X}_T$ over all possible positions. Hence for each random variable X_{tj}^τ in the component vectors X_t^τ of $\mathbf{X}_T^\tau$, we now designate the correlation $\rho(X_{tj}^\tau, Z_{\langle\tau+t\rangle j})$ between X_{tj}^τ and its corresponding representation $Z_{\langle\tau+t\rangle j}$ in $Z_{\langle\tau+t\rangle}$ as the (τ,t,j)-*positional correlation* between $\mathbf{X}_T$ and **Z**.

[3]To see this, recall that T-periodicity implies that $[Z_t,..,Z_{t+T}] = [Z_{\langle t\rangle},..,Z_{\langle t+T\rangle}]$ for all $t \in T$, and that each interval $[\langle t\rangle,..,\langle t+T\rangle]$ is by definition on the circle in Figure 2.1. 1

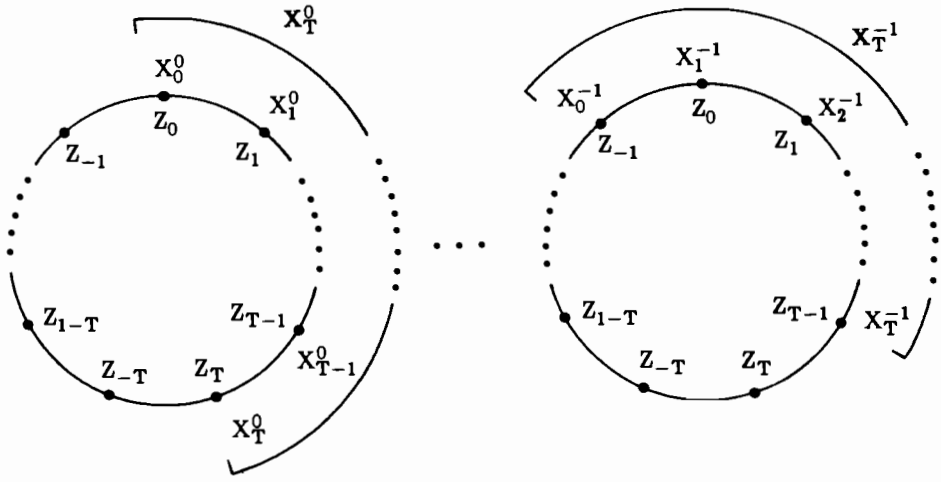


Figure 2.2 Positional Replications

Given these positional correlation measures, we seek finally to determine a T-stationary process $\mathbf{Z}^* \in \mathbf{P}_T$ which in some sense best represents $\mathbf{X}_T$ with respect to this entire set of criteria. To identify such a process $\mathbf{Z}^*$, observe first that it is naturally desirable for all these correlations with $\mathbf{Z}^*$ to be *positive*. Hence, if we now denote the *minimum positional correlation* between $\mathbf{X}_T$ and $\mathbf{Z} \in \mathbf{P}_T$ by

$$\rho(\mathbf{X}_T,\mathbf{Z}) = \min_{(\tau,t,j)} \rho(X^{\tau}_{tj}, Z_{\langle\tau+t\rangle j}) \tag{2.3}$$

then it is appropriate to focus our attention on the set

$$\mathbf{P}_T^+ = \{\mathbf{Z} \in \mathbf{P}_T \mid \rho(\mathbf{X}_T,\mathbf{Z}) > 0\} \tag{2.4}$$

of T-stationary processes which are *uniformly positively correlated* with $\mathbf{X}_T$. Moreover, it is evident from the definition of $\mathbf{P}_T^+$ that if there exists a process $\mathbf{Z}^* \in \mathbf{P}_T^+$ for which this minimum positional correlation is maximized, i.e. for which

$$\rho(\mathbf{X}_T,\mathbf{Z}^*) = \max_{\mathbf{Z}\in\mathbf{P}_T^+} \rho(\mathbf{X}_T,\mathbf{Z}) \tag{2.5}$$

then from among all T-stationary processes, $\mathbf{Z}^*$ may be said to exhibit maximal uniform agreement with $\mathbf{X}_T$ over all possible positions which this segment could occupy. Hence we now designate such a process $\mathbf{Z}^*$ as a *best uniform representative* of $\mathbf{X}_T$ in $\mathbf{P}_T^+$.

In this context, our main result (Theorem 4.1 below) is to show that there exists an essentially unique best uniform representative of $\mathbf{X}_T$ in $\mathbf{P}_T$. Moreover, this T-stationary process $\mathbf{C} = \{C_t \mid t \in \mathbf{T}\}$, which we now designate as a *circular-smoothing process*, is constructable from $\mathbf{X}_T$ by simply averaging the contributions of all positional replicates of $\mathbf{X}_T$ to each location on the circle in Figure 2.2. To be more explicit, observe from Figure 2.2 that the vector components of each positional replicate $\mathbf{X}_T^\tau$ which correspond to any given location t on the circle are precisely those in the t-th column of the matrix in Figure 2.3. Hence the finite representation $\mathbf{C}[T] = [C_{-T},..,C_0,..,C_T]$ of the desired T-periodic process $\mathbf{C}$ is constructable by simply averaging the random vectors in each column of this matrix. The resulting *circular smoothing* $\mathbf{C} = \{C_t \mid t \in \mathbf{T}\}$ of $\mathbf{X}_T$ is thus defined formally for all $t \in \mathbf{T}$ by:

$$C_t = \frac{1}{\sqrt{T+1}} \sum_{\tau=0}^{T} X_\tau^{\langle t-\tau \rangle}. \tag{2.6}$$

The choice of scaling factor $(1\sqrt{T+1})$ is designed to ensure that each random vector C_t in $\mathbf{C}$ has precisely the same *autocovariance structure* K(0) as the original random vectors X_t in $\mathbf{X}$ (as shown in expression (4.9) below).

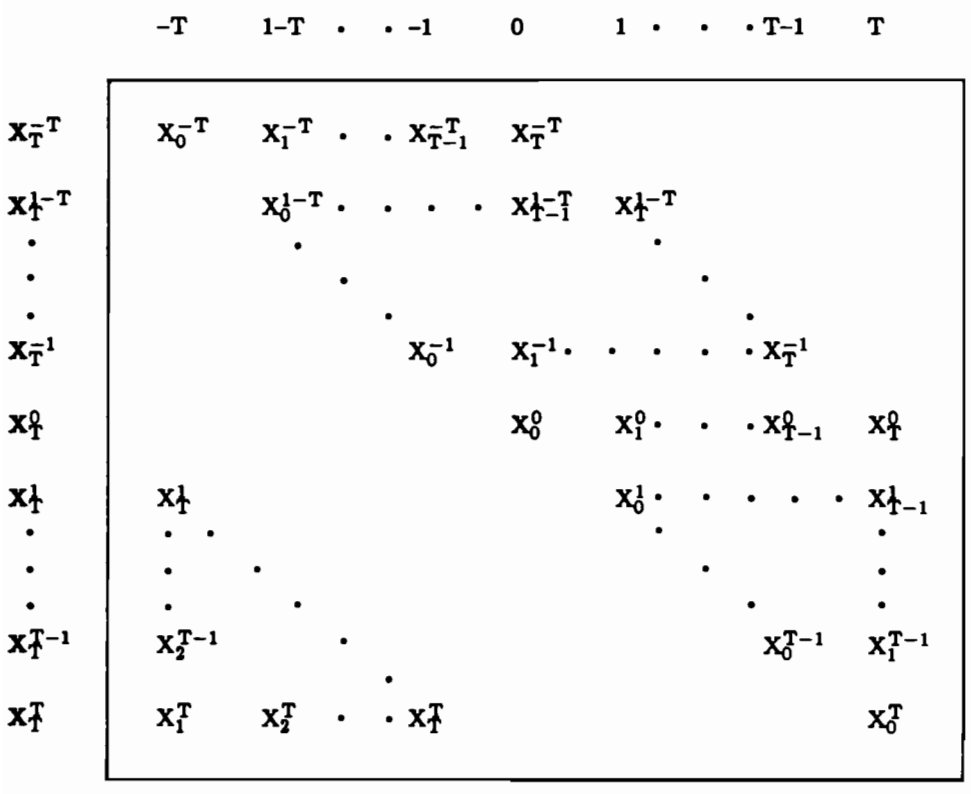


Figure 2.3 Contributions of Positional Replications

2.4 Spectral Representations of Circular-Smoothing Processes

From an analytical viewpoint, the central feature of circular-smoothing processes is the simplicity of their spectral representation. First of all, this representation is expressible in terms of ordinary random vectors, and is derivable from elementary considerations. Second, and even more important, the covariance 'spectrum' of this representation corresponds precisely to the *periodogram* of the original sample segment $\mathbf{X}_T = [X_0,..,X_T]$, and thus provides a new and simple theoretical interpretation of standard spectral estimators.

To state these results more precisely, we first consider the spectral representation of a general T-stationary process $\mathbf{Z}$. For simplicity, here we only consider the real-valued spectral representation of $\mathbf{Z}$ (the complex-valued spectral representation is derived in Section 3.1 below and compared both statistically and geometrically with the real-valued representation in Section 3.2 below). For this case, the appropriate *spectral representation theorem* (Theorem 3.1 below) shows that each random vector Z_t of $\mathbf{Z}$ can be decomposed into a trigonometric sum

$$Z_t = W_0 + \sum_{\tau=1}^{T} + [\cos(2\pi\tau t/2T+1)\, W_\tau + \sin(2\pi\tau t/2T+1)V_\tau] \tag{2.7}$$

of random *spectral-weight vectors* $(W_0,W_1,..,W_T,V_1,..,V_T)$. Moreover, these random vectors are 'linearly uncorrelated' in the sense that for every linear function $a'Z_t$ of Z_t, the corresponding random variables $(a'W_0,..,a'W_T,a'V_1,..,a'VT)$ are mutually uncorrelated (see Corollary 3.1 below). As mentioned in the introduction, this representation closely resembles the orthogonal decomposition of the finite representation $[Z_{-T},..,Z_0,..,Z_T]$ of $\mathbf{Z}$ into principle components. The relation between these two orthogonal decompositions is examined in Section 3.3 below, where it is shown that the spectral-weight vectors are proportional to an orthonormal transformation of the principle components. Hence they may be viewed in geometric terms as essentially a rotation of the principle components.

The primary advantage of this spectral representation over the more general principle components representation is the simple form of the trigonometric coefficients in (2.7). In particular, if for each $\tau = 1,..,T$ we now let $\omega_\tau = \tau/2T+1$, then $\cos(2\pi\omega_\tau t)$ and $\sin(2\pi\omega_\tau t)$ are both seen to be periodic functions of time with the same *frequency* ω_τ. Hence, in the spectral representation of $\mathbf{Z}$, each random vector Z_t is decomposed into a sum of uncorrelated *cyclical components* which may serve to reveal hidden periodicities in the stochastic structure of $\mathbf{Z}$. The geometric interpretation of these cyclical components is considered in more detail in Section 3.2 below.

Finally, it may be observed that this spectral representation is precisely the finite analogue of the more general infinite-dimensional representation of stationary processes by stochastic Fourier integrals (see the Remark following the statement of Theorem 3.1). In the present representation, the abstract notion of a stochastic integral is replaced by a simple sum of random variables, which is much more readily interpretable. A number of consequences of this interpretation are developed in detail in Smith (1981). For our present purposes, the single most important consequence of this finite representation is that the abstract 'cospectral' and 'quadrature spectral' moment-density matrices (associated, respectively, with the real and imaginary parts of the stochastic-integral representation) are here interpretable directly in terms of the covariances among spectral-weight vectors. In particular, it is shown in Section 4.3 below that the appropriate *cospectrum* $\Gamma^c_{\mathbf{Z}}$ and *quadrature spectrum* $\Gamma^q_{\mathbf{Z}}$ for $\mathbf{Z}$ are expressible as scaled covariance matrices defined, respectively, for each frequency ω_τ $(\tau = 1,..,T)$ as follows:

$$\Gamma^c_Z(\omega_\tau) = \frac{2T+1}{4\pi}\,\mathrm{cov}(W_\tau,W_\tau) = \frac{2T+1}{4\pi}\,\mathrm{cov}(V_\tau,V_\tau) \tag{2.8}$$

$$\Gamma^q_Z(\omega_\tau) = \frac{2T+1}{4\pi}\,\mathrm{cov}(W_\tau,V_\tau) = -\frac{2T+1}{4\pi}\,\mathrm{cov}(V_\tau,W_\tau)\,. \tag{2.9}$$

The pair of functions (Γ^c_Z, Γ^q_Z) is designed simply as the **Z-*spectrum*.**

Finally, turning to the specific case of *circular-smoothing processes* $\mathbf{C} = \{C_t \mid t \in T\}$, it is shown in Section 4.3 below that the C-spectrum corresponds precisely to the *periodogram* of the original sample segment X_T. More formally, if for a given sample segment $X_T = [X_0,..,X_T]$ we now define the *coperiodogram* P^c_T and *quadrature periodogram* P^q_T (corresponding to the real and imaginary parts of the complex-valued periodogram for X_T), respectively, for each frequency ω_τ (τ=1,..,T) by:

$$P^c_T(\omega_\tau) = \frac{1}{2\pi}\sum_{h\in[T]}\left(\frac{T+1-|h|}{T+1}\right)\cos(2\pi\,\omega_\tau\,h)\,K(h) \tag{2.10}$$

$$P^q_T(\omega_\tau) = -\frac{1}{2\pi}\sum_{h\in[T]}\left(\frac{T+1-|h|}{T+1}\right)\sin(2\pi\,\omega_\tau\,h)\,K(h) \tag{2.11}$$

then our central result is to show that

$$\Gamma^c_C(\omega_\tau) = P^c_T(\omega_\tau) \quad , \quad \tau=1,..,T \tag{2.12}$$

$$\Gamma^q_C(\omega_\tau) = P^q_T(\omega_\tau) \quad , \quad \tau=1,..,T\,. \tag{2.13}$$

In other words, if the pair of functions (P^c_T, P^q_T) is now designated as the X_T-*periodogram*, then (2.12) and (2.13) show that the X_T-periodogram is equivalent to the C-spectrum for the circular smoothing C of X_T. In view of this equivalence, the standard sample periodogram

$$\hat{P}^c_T(\omega_\tau) = \frac{1}{2\pi}\sum_{h\in[T]}\left(\frac{T+1-|h|}{T+1}\right)\cos(2\pi\,\omega_\tau\,h)\,\hat{K}(h) \tag{2.14}$$

$$\hat{P}^q_T(\omega_\tau) = -\frac{1}{2\pi}\sum_{h\in[T]}\left(\frac{T+1-|h|}{T+1}\right)\sin(2\pi\,\omega_\tau\,h)\,\hat{K}(h) \tag{2.15}$$

constructed in terms of the *sample covariance kernel* $\{\hat{K}(h) \mid h\in\in[T]\}$ for a given set of observations $[X_0,..,X_T]$ is now seen to yield a direct (unbiased) estimate of the C-spectrum.[4] Hence the standard periodogram techniques for estimating the infinite spectra

[4]The sample covariance kernel $\hat{K}$ for a stationary segment $X_T = [X_0,..,X_T]$ is defined by $\hat{K}(h) =$

of stationary processes **X** can now be given a new theoretical interpretation. In particular, the correspondence above shows that use of sample $\mathbf{X}_T$-periodograms to approximate the infinite spectrum for **X** is theoretically equivalent to direct estimation of the finite C-spectrum for $\mathbf{X}_T$.[5]

3. SPECTRAL REPRESENTATION OF FINITELY-STATIONARY PROCESSES

The spectral representation of finitely-stationary (circular) processes is well known for the univariate case (see for example Anderson (1971), Brillinger (1975), Fuller (1976), and Streitberg (1979)). However, the multivariate extensions of these results are far less well known (see Wahba (1968) and Fuller (1976; sector 7.4) for treatments of the bivariate case). Moreover, since these results are of central importance for finite spectral analysis, it is appropriate to present a self-contained development of the multivariate spectral representation theorem at this time.

To do so, let us begin by observing that if a given n-process $\mathbf{Z} = \{Z_t \mid t \in T\}$ is T-stationary, then the *covariance matrix* K for the finite representation $\mathbf{Z}[T] = [Z_{-T},..,Z_0..,Z_T]$ of **Z** must have the patterned structure shown in Figure 3.1.

This matrix is referred to as 'block circulant' because the individual covariance matrices [K(-T),..,K(0),.K(T)] are seen to be rotated by one position in each adjacent column (or row) block. Hence, if we observe that these column rotations can be accomplished by the *permutation matrix* P shown below,

$$P = \begin{bmatrix} 0\,1\,0\cdots 0 \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot\ 0 \\ 0 \quad\quad 1 \\ 1\,0\cdots 0 \end{bmatrix} \quad \Rightarrow \quad P \begin{bmatrix} -T \\ \cdot \\ \cdot \\ 0 \\ \cdot \\ T \end{bmatrix} = \begin{bmatrix} 1-T \\ \cdot \\ \cdot \\ 1 \\ \cdot \\ -T \end{bmatrix} \tag{3.1}$$

then it is clear that K can be written as an appropriate sum of successive permutations of the basic covariance matrices [K(-T),..K(0),..,K(T)]. To do so in a simple way, it is convenient to introduce tensor-product notation. For if we now define the *tensor product* A⊗B of any two matrices A and B by

---

$(T+1-h)^{-1}\ \Sigma_{t=0}^{T-h}\ X_{t+h}\ X_t'$ for h=0,...,T and $\hat{K}(h) = (T+1-h)^{-1}\ \Sigma_{t=-h}^{T}\ X_t\ X_{t+h}'$ for h=-1,..,-T. These statistics are easily seen to yield unbiased estimates of the covariance kernel K (see for example Fuller(1976; section 6.5)), so that (2.14) and (2.15) in turn yield unbiased estimates of the $\mathbf{X}_T$-periodogram.

[5]This theoretical equivalence also shows that the classical problem of constructing consistent estimates of infinite spectra has a finite interpretation which is directly analogous to the 'heteroscedasticity problem' for ordinary regression residuals (as observed by Duncan and Jones (1966) for example). For as in the case of variance estimation under general heteroscedasticity, the number of spectral moments to be estimated in the C-spectrum grows as fast as the sample size. Hence there is no hope of obtaining consistent estimators without further structural assumptions. As one possibility, if the C-spectrum is assumed to be 'sufficiently constant' over small frequency bands, then following Duncan and Jones (1966), one may construct consistent spectral estimates by pooling $\mathbf{X}_T$-periodogram values over these frequency bands. Alternatively, if the covariances between sample $\mathbf{X}_T$-periodogram matrices at different frequencies are assumed to be 'sufficiently small', then consistent estimators can also be constructed in terms of appropriate moving averages of these periodogram values (see for example Theorem 7.2.2. and Corollary 7.2.2. in Fuller (1976)).

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \tag{3.2}$$

then the circulant structure of the column blocks in K can be displayed explicitly as

$$K = P^{-T} \otimes K(-T) + \cdots + P^{T} \otimes K(T) = \sum_{\tau \in [T]} P^{\tau} \otimes K(\tau) \tag{3.3}$$

where each τ-th power P^{τ} of P corresponds to a column rotation by τ positions. The general class of block-circulant matrices representable by (3.3) is fully analyzed in Davis (1979; section 5.6). By applying these general results to the special case of covariance matrices (i.e. symmetric positive semidefinite matrices) we shall obtain a 'block diagonalization' of K which is the core of the spectral representation.

	Z_{-T}	Z_{1-T}	Z_{-1}	Z_0	Z_1	Z_{T-1}	Z_T
Z_{-T}	K(0)	K(1)	K(T-1)	K(T)	K(-T)	K(-2)	K(-1)
Z_{1-T}	K(-1)						K(-2)
Z_{-1}	K(1-T)		K(0)	K(1)			K(-T)
Z_0	K(-T)		K(-1)	K(0)	K(1)		K(T)
Z_1	K(T)			K(-1)	K(0)		K(T-1)
Z_{T-1}	K(2)						K(1)
Z_T	K(1)	K(2)	K(T)	K(-T)	K(1-T)	K(-1)	K(0)

Figure 3.1 Block Circulant Covariance Matrix

The key to this construction is the observation that all powers P^{τ} of the permutation matrix can be simultaneously diagonalized by a simple 'Fourier' matrix as follows. For each $\tau \in [T]$ let

$$r_{\tau} = \exp(i2\pi\tau/2T+1) = \cos(2\pi\tau/2T+1) + i \sin(2\pi\tau/2T+1) \tag{3.4}$$

where $i = \sqrt{-1}$ and observe that $r_\tau^{2T+1} = \cos(2\pi\tau) + i \sin(2\pi\tau) = 1$ for all τ. Hence (3.4) defines the distinct (2T + 1)-th *roots of unity*. If for each power $v \in [T]$ of the τ-th root we designate

$$f_{v\tau} = \frac{1}{\sqrt{2T+1}} r_\tau^v \quad , \quad \tau, v \in [T] \tag{3.5}$$

as the (v,τ)-th *Fourier coefficient*, then these coefficients define the following symmetric *Fourier matrix*:

$$F = \begin{bmatrix} f_{-T,-T} & \cdots & f_{-T,T} \\ \vdots & & \vdots \\ f_{T,-T} & \cdots & f_{T,T} \end{bmatrix} = \frac{1}{\sqrt{2T+1}} \begin{bmatrix} r_{-T}^{-T} & \cdots & r_T^{-T} \\ \vdots & & \vdots \\ r_{-T}^{T} & \cdots & r_T^{T} \end{bmatrix} . \tag{3.6}$$

Finally, recalling the definition of T-equivalents $\langle t \rangle$ for each integer $t \in \mathbf{T}$, and observing that $r_{-\tau} r_\tau^v = r_\tau^{v-1} = r_\tau^{\langle v-1 \rangle}$ holds identically for all $v \in \mathbf{T}$, it follows that multiplication of the τ-th column of F by $r_{-\tau}$ effectively rotates all elements up by one position. Hence letting

$$\Delta = \text{diag}(r_T, \ldots r_0, \ldots, r_T) \tag{3.7}$$

denote the diagonal matrix of roots of unity, we may conclude that

$$PF = F\Delta . \tag{3.8}$$

In other words, each root r_τ is a (complex) eigen value of the permutation matrix P with associated eigen vector equal to the $(-\tau)$-th column of F. Moreover, if for each complex number $c = a+ib$ we let $\bar{c} = a-ib$ denote the *conjugate* of c, and for each complex matrix $C = (c_{jk})$ let $\bar{C} = (\bar{c}_{jk})$ and $C^* = \bar{C}'$ denote the *conjugate* and *conjugate transpose* of C, respectively, then by employing the identity

$$\sum_{v \in [T]} r_\tau^v = \begin{cases} 2T + 1 & , \quad \tau = 0 \\ 0 & , \quad \tau \neq 0 \end{cases} \tag{3.9}$$

one may readily verify that $F^*F = FF^* = I$, and hence that $F^{-1} = F^*$. Thus F is a *unitary* matrix (the complex generalization of an orthonormal matrix), and (3.8) may be written as[6]

$$P = F\Delta F^* . \tag{3.10}$$

This diagonalization of P implies at once that

[6]This result is easily seen to be equivalent to Theorem 3.2.1 in Davis (1979) with appropriate relabeling of the elements of F and Δ.

$$P^{\tau} = (F\Delta F^*)(F\Delta F^*) \cdots (F\Delta F^*) = F\Delta^{\tau}F^* \tag{3.11}$$

so that all powers of P are, indeed, simultaneously diagonalized by F. This means that K may be further decomposed in terms of F by employing the tensor-product identities $AB\otimes CD = (A\otimes C)(B\otimes D)$ and $\sum_{\tau} A\otimes B_{\tau} = A\otimes\sum_{\tau} B_{\tau}$ to write K as

$$\begin{aligned} K &= \sum_{\tau\in[T]} (F\Delta^{\tau}F^*)\otimes(IK(\tau)I) \qquad (3.12)\\ &= \sum_{\tau\in[T]} (F\otimes I)\,[\Delta^{\tau}\otimes K(\tau)](F^*\otimes I) \\ &= (F\otimes I)[\sum_{\tau\in[T]} \Delta^{\tau}\otimes K(\tau)](F^*\otimes I)\,. \end{aligned}$$

This is in fact the block-diagonalization of K which we seek, as can be seen more clearly by noting that

$$\sum_{\tau\in[T]} \Delta^{\tau}\otimes K(\tau) = \begin{bmatrix} \sum_{\tau\in[T]} r_{T}^{\tau} K(\tau) & & \\ & \ddots & \\ & & \sum_{\tau\in[T]} r_{-T}^{\tau} K(\tau) \end{bmatrix} = \begin{bmatrix} H_{-T} & & \\ & \ddots & \\ & & H_{T} \end{bmatrix} \tag{3.13}$$

where for each $\tau\in[T]$

$$H_{\tau} = \sum_{v\in[T]} r_{-\tau}^{v} K(v) = \sum_{v\in[T]} \bar{r}_{\tau}^{v} K(v)\,. \tag{3.14}$$

Hence, employing this notation, and observing that $(F^*\otimes I) = (F\otimes I)^*$, we may rewrite (3.12) as

$$K = (F\otimes I) \begin{bmatrix} H_{-T} & & \\ & \ddots & \\ & & H_{T} \end{bmatrix} (F\otimes I)^*\,. \tag{3.15}$$

Finally, observing that $(F\otimes I)$ is itself unitary (since $(F\otimes I)(F\otimes I)^* = FF^*\otimes I = I$) we arrive at the fundamental result that: *every finitely-stationary covariance matrix is block-diagonalized by the same unitary matrix* $(F\otimes I)$.[7]

[7]This result is a special case of Theorem 5.6.4 in Davis (1979), as may be seen by writing expression (3.12) in the following alternative form:

3.1 Complex Spectral Representation

To utilize this decomposition, it is convenient to express the finite representation $[Z_{-T},..,Z_0,..,Z_T]$ of **Z** as a single vector

$$Z = \begin{bmatrix} Z_{-T} \\ \cdot \\ \cdot \\ Z_T \end{bmatrix} \quad \Rightarrow \quad \text{cov}(Z) = K . \tag{3.16}$$

With this notation, the following linear transformation

$$\frac{1}{\sqrt{2T+1}} (F \otimes I)^* Z = S = \begin{bmatrix} S_{-T} \\ \cdot \\ \cdot \\ S_T \end{bmatrix} \tag{3.17}$$

yields a complex-valued random vector S which we designate as the *complex spectral-weight vector*. If we express this random vector in terms of real and imaginary parts as $S = X + iY$, then the covariance matrix for S may be defined as

$$\begin{aligned} \text{cov}(S) &= E(SS^*) = E[(X + iY)(X'-iY')] \\ &= [E(XX')+E(YY')] + i[E(YX') - E(XY')] \end{aligned} \tag{3.18}$$

where E(XX'), E(YY'), E(XY'), and E(YX') are ordinary covariance matrices for the real random vectors X and Y (see for example Fuller (1976; section 1.5)). With this definition, it follows at once from (3.15) through (3.17) that

$$\text{cov}(S) = \frac{1}{2T+1} (F \otimes I)^* K (F \otimes I) = \frac{1}{2T+1} \begin{bmatrix} H_{-T} & & \\ & \ddots & \\ & & H_T \end{bmatrix} . \tag{3.19}$$

$$\begin{aligned} K &= (F \otimes I) \left[\sum_{\tau} (I \Delta^{\tau} I) \otimes (FF^* A_{\tau} FF^*)\right] (F \otimes I)^* \\ &= (F \otimes I)(I \otimes F) \left[\sum_{\tau} \Delta^{\tau} \otimes F^* A_{\tau} F\right] (I \otimes F^*)(F^* \otimes I) \\ &= (F \otimes F) \left[\sum_{\tau} \Delta^{\tau} \otimes F^* A_{\tau} F\right] (F \otimes F)^* . \end{aligned}$$

Hence, the appropriate relabeling of elements in F and Δ (mentioned in footnote 6 above) is easily seen to yield the special case of Theorem 5.6.4 in which n = m.

Hence the n-vector components $(S_{-T},..,S_0,..,S_T)$ of S are seen to be mutually *uncorrelated* random vectors with respective covariance matrices

$$\text{cov}(S_\tau) = \frac{1}{2T+1}\, H_\tau \quad , \quad \tau \in [T] \,. \tag{3.20}$$

Finally, observe from (3.17) together with the definition of Fourier coefficients in (3.5) that Z can be expressed as

$$\begin{bmatrix} Z_{-T} \\ \cdot \\ \cdot \\ Z_T \end{bmatrix} = \sqrt{2T+1}\,(F \otimes I)S = \begin{bmatrix} r_{-T}^{-T} I & \cdots & r_T^{-T} I \\ \cdot & & \cdot \\ r_{-T}^{T} I & \cdots & r_T^{T} I \end{bmatrix} \begin{bmatrix} S_{-T} \\ \cdot \\ \cdot \\ S_T \end{bmatrix} . \tag{3.21}$$

Thus, from (3.4) it follows that each random vector Z_t in **Z** can be written explicitly in terms of $(S_{-T},..,S_0,..,S_T)$ as

$$Z_t = \sum_{\tau \in [T]} r_\tau^t\, S_\tau = \sum_{\tau \in [T]} S_\tau \exp(i\,\lambda_\tau t) \quad , \quad t \in \mathbf{T} \tag{3.22}$$

where $\lambda_\tau = 2\pi\tau/2T+1$ for each $\tau \in [T]$. This representation of **Z** in terms of the uncorrelated complex random vectors $(S_{-T},..,S_0,..,S_T)$ is designated as the *complex spectral representation* of **Z**.

Each term in this representation is designated as the *complex cosinusoidal component* $C_\tau(t) = S_\tau \exp(i\lambda_\tau t)$ of **Z** with *angular frequency* λ_τ. In order to interpret these components in geometric terms, it is convenient to focus on a selected individual (univariate) process $\mathbf{Z}_j = \{Z_{tj} \mid t \in \mathbf{T}\}$ within the n-process **Z**. If we express the corresponding spectral weights $S_{\tau j} = X_{\tau j} + iY_{\tau j}$ in polar-coordinate form as $S_{\tau j} = |S_{\tau j}| \exp(iP_{\tau j})$, where $|S_{\tau j}| = (X_{\tau j}^2 + Y_{\tau j}^2)^{1/2}$ and $P_{\tau j} = \tan^{-1}(Y_{\tau j}/X_{\tau j})$, then each cosinusoidal component in the spectral representation of $\mathbf{Z}_j$ can be written as

$$\begin{aligned} C_{\tau j}(t) &= |S_{\tau j}| \exp\,[i(\lambda_\tau t + P_{\tau j})] \\ &= |S_{\tau j}| \cos(\lambda_\tau t + P_{\tau j}) + i\,|S_{\tau j}| \sin(\lambda_\tau t + P_{\tau j}) \,. \end{aligned} \tag{3.23}$$

Hence, if for any realization of the (real) random variables $|S_{\tau j}|$ and $P_{\tau j}$ we plot out the resulting function $C_{\tau j}(t)$ in the complex plane, as in Figure 3.2 (where each complex number c = x+iy is represented by the pair (x,y)), then in geometric terms this function is seen to wrap the real-time line around a circle of radius $|S_{\tau j}|$ about the origin. In other words, the passage of time is here represented geometrically by (counter-clockwise) rotation around this circle. In particular, the time coefficient λ_τ determines the arc angle (in radians) swept per unit of time, and is indeed seen to be the *angular frequency* of rotation. Similarly the time translation term $P_{\tau j}$ is seen to determine the position of time t = 0 on the circle, i.e. the initial *phase angle* of rotation, and the radius $|S_{\tau j}|$ is seen to determine the *amplitude* of rotation. For further discussion of these complex cosinusoidal

functions see, for example, Dhrymes (1970; sect. 10.1), Brillinger (1975; sect. 1.2), and Fuller (1976; sect. 3.2).

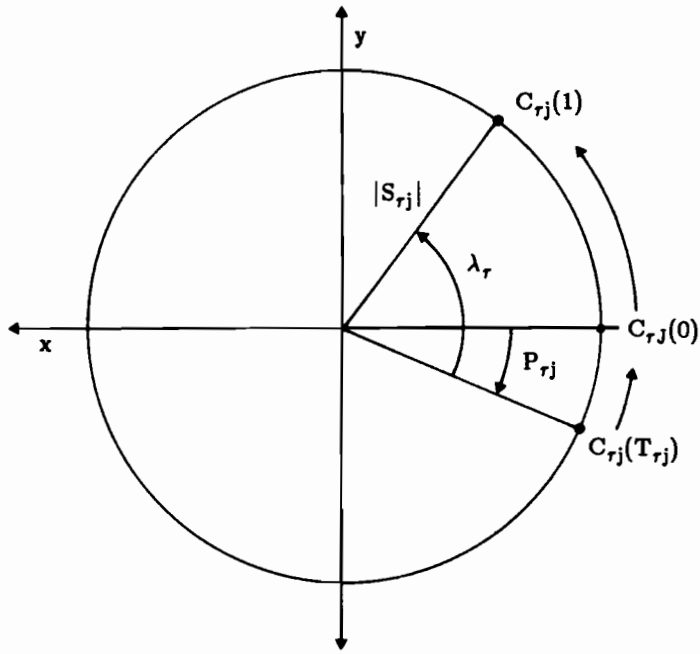


Figure 3.2 Complex Cosinusoidal Functions

3.2 Real Spectral Representation

The final step in our analysis of finitely-stationary processes is to transform the complex spectral representation above into real terms. To do so, we first observe that this representation can in fact be constructed entirely in terms of the subset of spectral-weight vectors $(S_0,S_1,..,S_T)$. In particular, the argument of (3.21) and (3.22) applied to (3.17) shows that

$$S_\tau = \frac{1}{\sqrt{2T+1}} \sum_v \bar{f}_{v\tau} Z_v = \frac{1}{2T+1} \sum_v \bar{r}^v_\tau Z_v$$

$$= \frac{1}{2T+1} \sum_v r^v_{-\tau} Z_v = \bar{S}_{-\tau} \qquad (3.24)$$

and hence that each random vector $S_{-\tau}$ is the *conjugate* of S_τ for $\tau = 1,...,T$. Employing this result, we may now transform the complex spectral representation of (3.22) into real terms as follows. If $\omega_\tau = \lambda_\tau/2\pi$ denotes the *time frequency* associated with each angular frequency λ_τ, then in terms of these real-time frequencies, we have the following real spectral representation of $\mathbf{Z}$ (where O denotes the zero matrix):

Theorem 3.1 *(Spectral Representation Theorem) For any real* T*-stationary* n*-process* $\mathbf{Z} = \{Z_t \mid t \in \mathbf{T}\}$ *there exist real random (spectral-weight) vectors* $(W_0, W_1, .., W_T, V_1, .., V_T)$ *such that for all* $t \in \mathbf{T}$,

$$Z_t = W_0 + \sum_{\tau=1}^{T} [\cos(2\pi\,\omega_\tau t) W_\tau + \sin(2\pi\,\omega_\tau t)\, V_\tau]. \tag{3.25}$$

In addition these random vectors satisfy

$$\operatorname{cov}(W_\tau, W_s) = \operatorname{cov}(V_\tau, V_s) = \operatorname{cov}(W_\tau, V_s) = O \tag{3.26}$$

for all $\tau = 0,1,..,T$ *and* $s = 1,..,T$ *with* $s \neq \tau$, *and satisfy*

$$\operatorname{cov}(W_\tau, W_\tau) = \operatorname{cov}(V_\tau, V_\tau) \tag{3.27}$$

$$\operatorname{cov}(W_\tau, V_\tau) = -\operatorname{cov}(V_\tau, W_\tau) \tag{3.28}$$

for all $\tau = 0,1,..,T$.

Remark: The statement of this theorem is seen to be exactly parallel to the more general infinite spectral representations of stationary vector processes. For example, expressions (3.25), (3.26), (3.27), and (3.28) are precisely the finite analogues of expressions (10.2.68), (10.2.69), (10.2.84), and (10.2.85) in Dhrymes (1970).

Proof: First, writing each complex spectral-weight vector as $S_\tau = X_\tau + i\,Y_\tau$, it follows from (3.5), (3.6) and (3.21) that each Z_t can be expressed as

$$Z_t = S_0 + \sum_{\tau=1}^{T} r_\tau^t S_\tau + \sum_{\tau=1}^{T} \bar{r}_\tau^t \bar{S}_\tau \tag{3.29}$$

$$= (X_0 + i\,Y_0) + \sum_{\tau=1}^{T} [\cos(2\pi\,\omega_\tau t) + i \sin(2\pi\,\omega_\tau t)]\,(X_\tau + i\,Y_\tau)$$

$$+ \sum_{\tau=1}^{T} [\cos(2\pi\,\omega_\tau t) - i \sin(2\pi\,\omega_\tau t)]\,(X_\tau - i\,Y_\tau)$$

$$= [X_0 + 2 \sum_{\tau=1}^{T} \cos(2\pi\,\omega_\tau t)\, X_\tau - 2 \sum_{\tau=1}^{T} \sin(2\pi\,\omega_\tau t)\, Y_\tau] + i\,Y_0 .$$

But since Z_t is real, we may conclude that $Y_0 = 0$. Hence, if we now define the *real spectral-weight vectors* $(W_0, W_1, .., W_T, V_1, .., V_T)$ by

$$W_0 = X_0 \tag{3.30}$$

$$W_\tau = 2X_\tau \quad , \quad \tau = 1,..,T \tag{3.31}$$

$$V_\tau = -2Y_\tau \quad , \quad \tau = 1,..,T \tag{3.32}$$

it follows at once from (3.29) that (3.25) must hold for this choice of random vectors. Thus it remains only to verify that conditions (3.26) through (3.28) hold for these vectors. To do so, observe first from (3.18) and (3.19) that for all $\tau \neq s$,

$$E(S_\tau S_s^*) = E(X_\tau X'_s + Y_\tau Y'_s) + i\, E(Y_\tau X'_s - X_\tau Y'_s) = O \tag{3.33}$$

which in turn implies that

$$E(X_\tau X'_s) = -E(Y_\tau Y'_s) \tag{3.34}$$

$$E(X_\tau Y'_s) = E(Y_\tau X'_s) \tag{3.35}$$

must both hold for all $\tau \neq s$. Next recall from (3.24) that $S_{-\tau} = \bar{S}_\tau \Rightarrow X_{-\tau} + i\, Y_{-\tau} = X_\tau - i\, Y_\tau$, and hence that both $X_{-\tau} = X_\tau$ and $Y_{-\tau} = -Y_\tau$ must hold identically for all τ. Thus it follows from (3.34) and (3.35) that for all $\tau \neq s$,

$$E(X_\tau X'_s) = E(X_{-\tau} X'_s) = -E(Y_{-\tau} Y'_s) = E(Y_\tau Y'_s) \tag{3.36}$$

$$E(X_\tau Y'_s) = E(X_{-\tau} Y'_s) = E(Y_{-\tau} X'_s) = -E(Y_\tau X'_s) \tag{3.37}$$

But (3.34) and (3.36) can only both hold if $E(X_\tau X'_s) = O = E(X_\tau Y'_s)$ for all $\tau \neq s$, and similarly, (3.35) and (3.37) can only both hold if $E(X_\tau Y'_s) = O$ for all $\tau \neq s$. Hence $E(X_\tau) = E(Y_\tau) = 0$ for all τ implies that

$$\text{cov}(X_\tau, X_s) = \text{cov}(Y_\tau, Y_s) = \text{cov}(X_\tau, Y_s) = O \quad , \quad \tau \neq s \tag{3.38}$$

and (3.26) follows at once from (3.38) together with $Y_0 = 0$ and the definitions in (3.30) through (3.32). Finally, to establish (3.27) and (3.28), observe that since $-\tau \neq \tau$ for all $\tau = 1,..,T$, the arguments of (3.36) and (3.37) also imply, respectively, that

$$E(X_\tau X'_\tau) = E(X_{-\tau} X'_\tau) = -E(Y_{-\tau} Y'_\tau) = E(Y_\tau Y'_\tau) \tag{3.39}$$

$$E(X_\tau Y'_\tau) = E(X_{-\tau} Y'_\tau) = E(Y_{-\tau} X'_\tau) = -E(Y_\tau X'_\tau) \tag{3.40}$$

for all $\tau = 1,..,T$. Hence by again employing the argument above (3.38) , we may conclude that (3.27) and (3.28) follow directly from (3.39) and (3.40) together with (3.31) and (3.32). End of proof.

Thus the *real spectral representation of* **Z** in (3.25) decomposes Z_t into a sum of uncorrelated *real cosinusoidal components* $R_\tau(t) = W_\tau \cos(2\pi\, \omega_\tau t) + V_\tau \sin(2\pi\, \omega_\tau t)$. As with the complex spectral representation in (3.22), these components may be interpreted geometrically by focusing on an individual univariate process $\mathbf{Z}_j$ within **Z** and writing the corresponding pairs of spectral weights $(W_{\tau j}, V_{\tau j})$ in polar-coordinate form as $W_{\tau j} =$

$A_{\tau j} \cos(2\pi\, \omega_\tau\, T_{\tau j})$ and $V_{\tau j} = A_{\tau j} \sin(2\pi\, \omega_\tau\, T_{\tau j})$, where $A_{\tau j} = (W_{\tau j}^2 + V_{\tau j}^2)^{1/2}$ and $T_{\tau j} = (1/2\pi\, \omega_\tau) \tan^{-1}(V_{\tau j}/W_{\tau j})$. In these terms, each real cosinusoidal component $R_{\tau j}(t)$ can be written as

$$\begin{aligned} R_{\tau j}(t) &= A_{\tau j} [\cos(2\pi\, \omega_\tau\, T_{\tau j}) \cos(2\pi\, \omega_\tau t) + \sin(2\pi\, \omega_\tau\, T_{\tau j}) \sin(2\pi\, \omega_\tau t)] \\ &= A_{\tau j} \cos[2\pi\, \omega_\tau\, (t - T_{\tau j})] \,. \end{aligned} \tag{3.41}$$

Hence, if for any realization of the random variables $A_{\tau j}$ and $T_{\tau j}$ we plot the resulting cyclical function $R_{\tau j}(t)$ as in Figure 3.3, then the parameters are seen to have a clear geometrical interpretation. First, since the duration or *period* of each cycle is given by $1/\omega_\tau$, it follows at once that ω_τ is indeed the *time frequency* of each cycle (i.e. cycles per unit of time). Similarly, $A_{\tau j}$ denotes the *amplitude* of each cycle, and $T_{\tau j}$ denotes the *phase shift* relative to the origin (i.e. the time interval from $t = 0$ to the nearest cycle peak). Hence, as in the complex case, each individual process $\mathbf{Z}_j$ in $\mathbf{Z}$ is seen to be representable in geometric terms as a sum of uncorrelated real cosinusoidal components $R_{\tau j}(t)$ with fixed time frequencies ω_τ but random amplitudes $A_{\tau j}$ and phase shifts $T_{\tau j}$. For further discussion of this representation see, for example, Fuller (1976; sect. 1.6) and Smith (1981; sects. 4.3 and 5.4).

Finally, to see the geometric equivalence between the real and complex spectral representation, observe first that $\lambda_\tau = 2\pi\, \omega_\tau$ by definition, and moreover from (3.31) and (3.32), that $-P_{\tau j} = 2\pi\, \omega_\tau\, T_{\tau j} = \lambda_\tau\, T_{\tau j}$. Hence λ_τ and $P_{\tau j}$ are simply the angular equivalents of ω_τ and $T_{\tau j}$, respectively. (For example, the time lapse $T_{\tau j}$ from $R_{\tau j}(0)$ to the first peak $R_{\tau j}(T_{\tau j})$ in Figure 3.3 corresponds precisely to the counter-clockwise rotation of $P_{\tau j}$ radians from $C_{\tau j}(0)$ to $C_{\tau j}(T_{\tau j})$ on the circle in Figure 3.2). In addition, (3.31) and (3.32) also imply that $2|S_{\tau j}| = A_{\tau j}$, and hence that the amplitude of $C_{\tau j}(t)$ is always proportional to the amplitude of $R_{\tau j}(t)$.

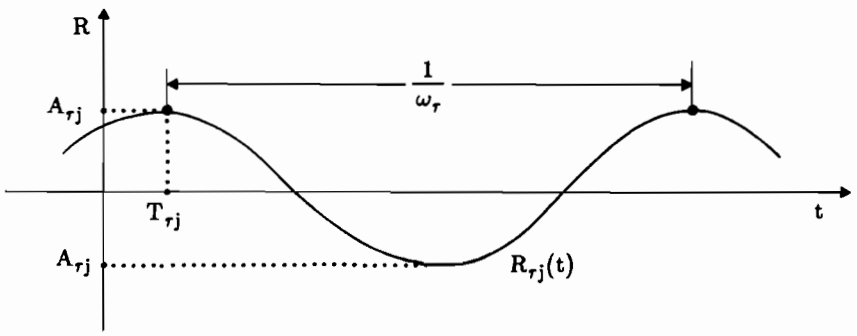


Figure 3.3 Real Cosinusoidal Functions

However, there remains one statistical difference between these two representations. For while the complex spectral-weight vectors $(S_{-T},..,S_0,..,S_T)$ are always uncorrelated, it follows from expression (3.28) of Theorem 3.1 that the pairs of real spectral-weight vectors (W_τ,V_τ) may indeed be correlated. With this in mind, it is of interest to observe that real spectral-weight vectors $(W_0,W_1,..,W_T,V_1,..,V_T)$ may nonetheless be regarded as *linearly uncorrelated* in the following sense:

> **Corollary 3.1** *The spectral-weight vectors* $(W_0,W_1,..,W_T,V_1,..,V_T)$ *in* (3.25) *satisfy the condition that for all fixed n-vectors* $a = (a_1,..,a_n)$ *the corresponding random variables* $(a'W_0,a'W_1,..,a'W_T,a'V_1,..,a'V_T)$ *are mutually uncorrelated.*

Proof: In view of (3.26) we need only consider the pairs of random variables $(a'W_\tau,a'V_\tau)$ for each $\tau = 1,..,T$. But since cov (X,Y) = cov(Y,X) holds identically for all random variables X and Y, and since (3.28) implies

$$\begin{aligned} \text{cov}(a'W_\tau,a'V_\tau) &= a'\text{cov}(W_\tau,V_\tau)\, a = -\, a'\text{cov}(V_\tau,W_\tau)\, a \\ &= -\, \text{cov}(a'V_\tau,a'W_\tau) \end{aligned} \tag{3.42}$$

it follows that $\text{cov}(a'W_\tau,a'V_\tau) = 0$ must hold identically for all linear compounds a and all $\tau = 1,..,T$. End of proof.

Hence, each univariate process $\mathbf{Z}_a = \{a'\mathbf{Z}_t \mid t \in \mathbf{T}\}$ defined as a linear function of a finitely-stationary process **Z** must have a spectral representation consisting of *uncorrelated* random spectral-weight variables. In particular, each individual component process $\mathbf{Z}_j$ of **Z** thus satisfies this condition.

3.3 Relation of Spectral Weights to Principle Components

Before leaving the general analysis of finitely-stationary processes, it is of interest to observe that each such process has a natural principle-component representation as well. In particular, if we employ the vector representation Z of such processes, as defined in expression (3.16), then we may construct such a representation by simply diagonalizing the covariance matrix K for Z. To do so, recall that every covariance matrix K can be written as $K = U\Delta U'$, where Δ is the nonnegative diagonal matrix of eigen values for K and U is the associated orthonormal matrix of eigen vectors (i.e. with $U'U = UU' = I$). The random vector of *principle components* P for Z is then defined by the linear transformation

$$P = U'Z\,. \tag{3.43}$$

By construction, P consists of mutually uncorrelated random variables since $\text{cov}(P) = U'KU = \Delta$ (for a more detailed analysis of these principle components, see for example Dhrymes (1970; sect. 2.2)).

The relation between these principle components and the complex spectral weights S defined in (3.17) is seen immediately by premultiplying (3.43) by U and substituting into (3.17) to obtain

$$S = \frac{1}{\sqrt{2T+1}} (F \otimes I)^* UP \ . \tag{3.44}$$

Since $(F \otimes I)^*$ is unitary and U is orthonormal, it then follows at once that $(F \otimes I)^* U$ is also unitary, i.e. it is a *rigid motion* in geometric terms.[8] Thus, to construct the complex spectral weights S from the principle components P, one simply rotates P into the complex plane by $(F \otimes I)^* U$, and scales the resulting vector by $1/\sqrt{2T+1}$.

A similar interpretation holds for the real spectral weights. To see this, observe that if we now let $W' = (W_0, W_1, .., W_T, V_1, .., V_T)$ denote the vector of all real spectral weights and express the set of trigonometric coefficients in (3.25) in matrix form as

$$R = \begin{bmatrix} 1 & \cos[2\pi\omega_1(-T)] & \ldots & \cos[2\pi\omega_T(-T)] & \sin[2\pi\omega_1(-T)] & \ldots & \sin[2\pi\omega_T(-T)] \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & \cos[2\pi\omega_1(T)] & \ldots & \cos[2\pi\omega_T(T)] & \sin[2\pi\omega_1(T)] & \ldots & \sin[2\pi\omega_T(T)] \end{bmatrix} \tag{3.45}$$

then (3.28) can be written simply as

$$Z = (R \otimes I)W \ . \tag{3.46}$$

Next, defining the positive diagonal matrix $D = (\sqrt{2T+1}, \sqrt{2T+1/2}, .., \sqrt{2T+1/2})$ and setting $M = RD^{-1}$, we may rewrite (3.46) as

$$Z = (MD \otimes I)W = (M \otimes I)(D \otimes I)W \ . \tag{3.47}$$

Finally, since M is by construction orthonormal,[9] it follows from (3.43) together with the tensor-product identity $(A \otimes B)^{-1} = (A^{-1}) \otimes (B^{-1})$ that W can be expressed in terms of P as

$$W = (D^{-1} \otimes I)(M' \otimes I)UP \ . \tag{3.48}$$

Hence, observing that $(M' \otimes I)U$ is an orthonormal matrix and that $(D^{-1} \otimes I)$ is a positive diagonal matrix, it again follows that the spectral weights W may be constructed from the principle components P by first rotating P with $(M' \otimes I)U$ and then scaling the resulting vector by $(D^{-1} \otimes I)$.

For the univariate case, it is worth noting that the rotation operation is unnecessary. To see this, observe simply that (3.47) reduces to Z = MDW in the univariate case. Moreover, since the spectral weights are in this case mutually uncorrelated (by Corollary 3.1), we must have $K = \text{cov}(Z) = MD\ \text{cov}(W)DM' = M\Delta M'$, where $\Delta = D\ \text{cov}(W)D$ is a

[8] Since $\|Ux\| = (x^*U^*Ux)^{1/2} = (x^*x)^{1/2} = \|x\|$ for every complex vector x under a unitary transformation U, it follows that all lengths of vectors and angles between vectors are preserved, and hence that U is the matrix representation of a *rigid motion* in the given vector space.

[9] With appropriate relabeling of the elements in the vector W and re-indexing of time, the matrix M is seen to be precisely the orthonormal matrix in expression (4.2.11) of Fuller (1976).

nonnegative diagonal matrix. Hence K is diagonalized by the orthonormal matrix M, and we may choose U = M in (3.43) without loss of generality. This means that (3.48) is reduced to $W = D^{-1}P$, and thus that the spectral weights in the univariate case are simply a scaled version of the principle components.

4. CIRCULAR-SMOOTHING PROCESSES

In this final section, the circular smoothing **C** of a stationary segment $\mathbf{X}_T = [X_0,..,X_T]$ is developed formally, and is shown to yield the best uniform representative of $\mathbf{X}_T$ from among all finitely-stationary processes. The results of Section 2 above are then employed to show that the spectral representation of **C** yields a new interpretation of the sample periodogram for $\mathbf{X}_T$.

To do so, it is necessary to begin by developing the probabilistic underpinnings for such processes in more detail. Given a set Ω together with a probability measure p over a σ-field **F** of subsets of Ω, the triple $\langle\Omega,\mathbf{F},p\rangle$ defines a fixed *probability space* over Ω. If **H** denotes the linear (Hilbert) space of random variables over $\langle\Omega,\mathbf{F},p\rangle$ with $E(X) = 0$ and $E(X^2) < \infty$, then each element $X = (X_1,..,X_n)$ of the n-fold product space $\mathbf{H}^n$ defines a *random* n-*vector* over $\langle\Omega,\mathbf{F},p\rangle$, and each sequence $\mathbf{X} = \{X_t \mid t\in\mathbf{T}\}$ in $\mathbf{H}^n$ defines a n-*process* on $\mathbf{H}^n$. If **P** denotes the class of n-processes on $\mathbf{H}^n$, then each finite segment $[X_t,..,X_{t+T}]$ of a stationary n-process $\mathbf{X} \in \mathbf{P}$ (i.e. satisfying (2.2) above) is designated as a *stationary segment* on $\mathbf{H}^n$. Our only assumption regarding the underlying probability space $\langle\Omega,\mathbf{F},p\rangle$ is that it is rich enough to allow the construction in $\mathbf{H}^n$ of independently and identically distributed *replicates* $\mathbf{X}_T^\tau = [X_0^\tau,..,X_T^\tau]$, $\tau\in[T]$, of any given stationary segment $\mathbf{X}_T = [X_0,..,X_T]$ on $\mathbf{H}^n$.[10] In terms of such replicates, we may then construct the *circular smoothing* $\mathbf{C} = \{C_t \mid t\in\mathbf{T}\}$ of $\mathbf{X}_T$ to be the n-process defined for all $t\in\mathbf{T}$ by

$$C_t = \frac{1}{\sqrt{T+1}} \sum_{\tau=0}^{T} X_\tau^{\langle t-\tau\rangle} \tag{4.1}$$

4.1 Finite Stationarity of Circular-Smoothing Processes

If, as in Section 2.2, we let $\mathbf{P}_T$ denote the class of T-*stationary processes* $\mathbf{Z} = \{Z_T \mid t\in\mathbf{T}\}$ in **P** (i.e. satisfying (2.2) and $Z_T = Z_{\langle t\rangle}$ for all $t\in\mathbf{T}$), then our present objective is to show that for any stationary segment $\mathbf{X}_T$ on $\mathbf{H}^n$, the associated circular smoothing **C** of $\mathbf{X}_T$ is T-stationary, i.e. that $\mathbf{C} \in \mathbf{P}_T$. To do so, it is useful to record the following properties of T-equivalents. First, observe that by definition $\langle t\rangle = t \Leftrightarrow |t| \leq T$. In addition,

$$\langle t+\tau\rangle = \langle t+\langle\tau\rangle\rangle = \langle\langle t\rangle + \langle\tau\rangle\rangle \tag{4.2}$$

$$\langle t+\tau\rangle = \langle s+\tau\rangle \Rightarrow \langle t\rangle = \langle s\rangle \tag{4.3}$$

---

[10] This may always be accomplished, for example, by replacing $\langle\Omega,\mathbf{F},p\rangle$ with the cartesian product space $\pi_{\tau\in[T]}\langle\Omega,\mathbf{F},p\rangle_\tau$ constructed from replicates $\langle\Omega,\mathbf{F},p\rangle_\tau$ of $\langle\Omega,\mathbf{F},p\rangle$, as for example in Halmos (1950; sec. 49).

hold for all t, s, $\tau \in \mathbf{T}$. Finally, for all $t,s \in [T]$,[11]

$$\langle t+s \rangle = 0 \quad \Rightarrow \quad t+s = 0 \,. \tag{4.4}$$

With these identities, it can now be shown that:

Lemma 4.1 *The circular smoothing* $\mathbf{C} = \{C_t \mid t \in \mathbf{T}\}$ *of a stationary segment* $\mathbf{X}_T = [X_0,..,X_T]$ *is* T-*stationary.*

Proof: First observe from (4.2) that since $\langle t-\tau \rangle = \langle\langle t \rangle - \tau \rangle$ holds identically for all t, $\tau \in \mathbf{T}$, it follows at once from the definition of **C** in (4.1) that $C_t = C_{\langle t \rangle}$ for all $t \in \mathbf{T}$, and hence that **C** is T-periodic. To verify that **C** is also stationary, it suffices to show that the covariances $\text{cov}(C_{t+\tau}, C_t)$ can be written solely as a function of τ. To do so, observe first that for any two replicates $[X_0^\tau,..,X_t^\tau]$ and $[X_0^v,..,X_t^v]$ of a stationary segment $[X_0,..,X_T]$ from a process with covariance kernel K, it follows from the independence of replicates that for all t,s = 0,1,..,T,

$$\text{cov}(X_t^\tau, X_s^v) = \begin{cases} K(t-s) & , \quad \tau = v \\ O & , \quad \tau \neq v \end{cases} . \tag{4.5}$$

Hence if we expand $\text{cov}(C_{t+\tau}, C_t)$ in terms of (4.1) as

$$\text{cov}(C_{t+\tau}, C_t) = \frac{1}{T+1} \sum_{v=0}^{T} \sum_{s=0}^{T} \text{cov}\,(X_v^{\langle t+\tau-v \rangle}, X_s^{\langle t-s \rangle}) \tag{4.6}$$

then it follows from (4.5) that the only nonzero terms in (4.6) will be those with $\langle t+\tau-v \rangle = \langle t-s \rangle$. Moreover, since $v,s \in [0,..,T]$ implies $|\tau - v| \leq T$, it also follows from (4.2) through (4.4) that

$$\begin{aligned} \langle t+\tau-v \rangle = \langle t-s \rangle \; &\Rightarrow \langle \tau - v \rangle = \langle -s \rangle \\ &\Rightarrow 0 = \langle 0 \rangle = \langle\langle t-v \rangle - \langle -s \rangle\rangle = \langle\langle \tau \rangle + s-v \rangle \\ &\Rightarrow v-s = \langle \tau \rangle \,. \end{aligned} \tag{4.7}$$

Hence, letting $S\langle\tau\rangle = \{s \mid s, s+\langle\tau\rangle \in [0,..,T]\}$, we can write (4.6) as

$$\begin{aligned} \text{cov}\,(C_{t+\tau}, C_t) &= \frac{1}{T+1} \sum_{s \in S\langle\tau\rangle} \text{cov}(X_{s+\langle\tau\rangle}^{\langle t-s \rangle}, X_s^{\langle t-s \rangle}) \\ &= \frac{1}{T+1} \sum_{s \in S\langle\tau\rangle} K(s + \langle\tau\rangle - s) \\ &= \frac{1}{T+1} \,(\#\, S\langle\tau\rangle)\, K\,(\langle\tau\rangle) \end{aligned} \tag{4.8}$$

[11] To see this, observe that $t+s - \langle t+s \rangle = n(2T+1)$ for some $n \in \mathbf{T}$, and hence that $\langle t+s \rangle = 0 \Rightarrow n\,(2T+1) = t+s \Rightarrow |n|\,(2T+1) = |t+s| \leq |t| + |s| = 2T$. Thus, n = 0 is the only possible solution for n, and (4.4) follows.

where # $S\langle\tau\rangle$ denotes the cardinality of $S\langle\tau\rangle$. Finally, to evaluate # $S\langle\tau\rangle$ explicitly, observe that since $s,s+\langle\tau\rangle \in [0,..T] \Rightarrow \max(0,\langle\tau\rangle) \leq s \leq \min(T,T+\langle\tau\rangle)$, it follows on the one hand that $\langle\tau\rangle \leq 0 \Rightarrow 0 \leq s \leq T - |\langle\tau\rangle| \Rightarrow \# S\langle\tau\rangle = T + 1 - |\langle\tau\rangle|$, and on the other hand, that $\langle\tau\rangle \geq 0 \Rightarrow \langle\tau\rangle \leq s \leq T \Rightarrow \# S\langle\tau\rangle = T + 1 - |\langle\tau\rangle|$. Hence in all cases we obtain $\# S\langle\tau\rangle = T + 1 - |\langle\tau\rangle|$, and (4.8) becomes

$$\operatorname{cov}(C_{t+\tau},C_t) = \left(\frac{T + 1 - |\langle\tau\rangle|}{T+1}\right) K(\langle\tau\rangle) . \tag{4.9}$$

Thus $\operatorname{cov}(C_{t+\tau},C_t)$ is seen to depend only on the time lag τ, and C is therefore T-stationary. End of proof.

Notice also that (4.9) yields an explicit relationship between the covariance kernels of C and $\mathbf{X}_T$. This relation will play a fundamental role in interpreting the spectral representation of C in section 4.3 below.

4.2 Best Uniform Representatives

Having established that C is finitely-stationary, we next show that C is *uniformly positively correlated* with $\mathbf{X}_T$, as defined in (2.3) and (2.4) above by

$$0 < \rho(\mathbf{X}_T,C) = \min_{(\tau,t,j)} \rho(X^{\tau}_{tj}, C_{\langle\tau+t\rangle j}) . \tag{4.10}$$

To do so, observe first that if e_j denotes the j-th column in the n-square identity matrix I, then (4.9) implies that $\operatorname{var}(C_{\langle\tau+t\rangle j}) = \operatorname{var}(e_j' C_{\langle\tau+t\rangle}) = e_j' \operatorname{cov}(C_{\langle\tau+t\rangle}) e_j = e_j' K(0) e_j = \operatorname{var}(X^{\tau}_{tj}) = \sigma_j^2$, so that by (4.1)

$$\begin{aligned} \rho(X^{\tau}_{tj}, C_{\langle\tau+t\rangle j}) &= \frac{1}{\sigma_j^2} \operatorname{cov}(e_j' X^{\tau}_t, e_j' C_{\langle\tau+t\rangle}) \\ &= \frac{1}{\sigma_j^2} \left[\frac{1}{\sqrt{T+1}} \sum_{s=0}^{T} \operatorname{cov}(X^{\tau}_t, X^{\langle\tau+t-s\rangle}_s)\right] . \end{aligned} \tag{4.11}$$

But from (4.6), these covariance matrices are all identically zero, except for the term with $\tau = \langle\tau+t-s\rangle$; and for this term it follows that $0 = \langle\tau-\langle\tau+t-s\rangle\rangle = \langle\tau-(\tau+t-s)\rangle = \langle s-t\rangle \Rightarrow s = t$. Hence

$$\rho(X^{\tau}_{tj}, C_{\langle\tau+t\rangle j}) = \frac{1}{\sigma_j^2} e_j' \left[\frac{1}{\sqrt{T+1}} K(0)\right] e_j = \frac{1}{\sqrt{T+1}} \frac{e_j' K(0) e_j}{\sigma_j^2} = \frac{1}{\sqrt{T+1}} \tag{4.12}$$

holds identically for all (τ,t,j), and we may conclude that

$$\rho(\mathbf{X}_T,\mathbf{C}) = \frac{1}{\sqrt{T+1}} > 0 \ . \tag{4.13}$$

Thus **C** is seen to be an element of the class $\mathbf{P}_T^+$ of finitely-stationary processes which are uniformly positively correlated with $\mathbf{X}_T$ (as defined in (2.4)).

To establish that **C** is a *best uniform representative* of $\mathbf{X}_T$ within this class, it remains to show that

$$\rho(\mathbf{X}_T,\mathbf{C}) = \max_{\mathbf{Z}\in\mathbf{P}_T^+} \rho(\mathbf{X}_T,\mathbf{Z}) \ . \tag{4.14}$$

With this objective in mind, we first require a number of preliminary results. For any finite set of random variables $\mathbf{S} = \{X_0,..,X_T\}$ in the Hilbert space **H** of zero-mean finite-variance random variables over $\langle\Omega,\mathbf{F},p\rangle$, let

$$\text{span}(\mathbf{S}) = \{X = \sum_{t=0}^{T} a_t X_t \mid a_0,..,a_T \in \mathbf{R}\} \tag{4.15}$$

denote the *linear span* of **S** in **H**, and for any random variable $Z \in \mathbf{H}$ let

$$\rho(Z,\mathbf{S}) = \min_{x\in\mathbf{S}} \rho(Z,X) \tag{4.16}$$

denote the *minimum correlation* between Z and **S**. Our first result is then to show that $\rho(Z,\mathbf{S})$ can only achieve a positive maximum among the random variables Z in the linear span of **S**. To establish this result, observe that the appropriate *inner product* between pairs of elements X and Y in the Hilbert space **H** is defined by their covariance cov(X,Y) = E(XY) as random variables. Hence, if we now let comp(**S**) = {$Y \in \mathbf{H} \mid X \in \text{span}(\mathbf{S}) \Rightarrow E(XY) = 0$} denote the *orthogonal complement* of the finite-dimensional subspace in span (**S**) in **H**, then it is well known (see for example Bachman and Narici (1966; theorem 1.8) that for each $Z \in \mathbf{H}$ there exist *unique* random variables $X \in \text{span}(\mathbf{S})$ and $Y \in \text{comp}(\mathbf{S})$ such that Z = X + Y. Moreover, if $E(Z^2) > 0$, then $Z \notin \text{span}(\mathbf{S}) \Rightarrow E(Y^2) > 0$. Hence, if we now designate this pair of random variables (X,Y) as the **S**-*representation* of Z then:

Lemma 4.2 *For any finite subset* $\mathbf{S} = \{X_0,..,X_T\} \subset \mathbf{H}$ *and any* $Z \in \mathbf{H}$ *with* **S**-*representation* (X,Y), *if* $Z \notin$ *span* (**S**) *then*

$$\rho(Z,\mathbf{S}) > 0 \Rightarrow \rho(Z,\mathbf{S}) < \rho(X,\mathbf{S}). \tag{4.17}$$

Proof: First observe that $Y \in \text{comp}(\mathbf{S})$ and $\{X,X_0,..,X_T\} \subset \text{span}(\mathbf{S})$ imply $E(XY) = E(X_tY) = 0$ for all t = 0,..,T. Moreover, since $Z \notin \text{span}(\mathbf{S}) \Rightarrow E(Y^2) > 0$, it then follows that for all t = 0,..,T, $\rho(Z,\mathbf{S}) > 0$ implies

$$\begin{aligned} 0 < \rho(Z,X_t) &= \rho(X+Y,X_t) \\ &= E(XX_t + YX_t)/\sigma(X_t)\sigma(X+Y) \\ &= E(XX_t)/\sigma(X_t)\,[E(X^2) + E(Y^2)]^{1/2} \\ &< E(XX_t)/\sigma(X_t)\sigma(X) \\ &= \rho(X,X_t) \ . \end{aligned} \tag{4.18}$$

Hence (4.17) follows at once from (4.18) together with the finiteness of **S**. End of proof.

In addition we require the following elementary inequality for positive numbers

Lemma 4.3 *For all positive numbers* $a_0,..,a_T$,

$$\min_t \frac{a_t}{\sqrt{\sum_s a_s^2}} \leq \frac{1}{\sqrt{T+1}} \tag{4.19}$$

with equality holding if $a_0 = \cdots = a_T$.

By employing these results, we can now establish the fundamental lemma upon which our main theorem depends. If each random variable X with $E(X) = 0$ and $E(X^2) = \sigma^2$ is now designated as a $(0,\sigma^2)$-random variable, then:

Lemma 4.4 *For any finite set of uncorrelated* $(0,\sigma^2)$*-random variables* $\mathbf{S} = \{X_0,..,X_T\} \subset \mathbf{H}$, *the unique* $(0,\sigma^2)$*-random variable* $X^* \in \mathbf{H}$ *satisfying*

$$\rho(X^*,\mathbf{S}) = \max_{X \in \mathbf{S}} \rho(X,\mathbf{S}) \tag{4.20}$$

is given by

$$X^* = \frac{1}{\sqrt{T+1}} \sum_{t=0}^{T} X_T . \tag{4.21}$$

Proof: First observe that since the X_t's are uncorrelated by hypothesis, it follows that $\sigma^2(X^*) = (T+1)^{-1} \Sigma_t E(X_t^2) = \sigma^2$ and $\rho(X^*,X_t) = 1/\sqrt{T+1}$ for all $t = 0,..,T$. Hence X^* is by construction a $(0,\sigma^2)$-random variable in span (**S**) with $\rho(X^*,\mathbf{S}) = 1/\sqrt{T+1} > 0$. Thus, it suffices to show that X^* is the unique $(0,\sigma^2)$-random variable in span (**S**) satisfying

$$\rho(X^*,\mathbf{S}) = \max_{X \in \text{span}(\mathbf{S})} \rho(X,\mathbf{S}) . \tag{4.22}$$

For if (X,Y) denotes the **S**-representation of any other random variable $Z \in \mathbf{H}$ satisfying $\rho(Z,\mathbf{S}) > 0$, then by Lemma 4.2 it will follow from (4.22) that $\rho(X^*,\mathbf{S}) \geq \rho(X,\mathbf{S}) > \rho(Z,\mathbf{S}) > 0$. To establish (4.22), suppose to the contrary that $\rho(X,\mathbf{S}) \geq \rho(X^*,\mathbf{S})$ for some $X \in \text{span}(\mathbf{S})$ with $X \neq X^*$. Then $X \in \text{span}(\mathbf{S})$ implies that $X = \Sigma_t a_t X_t$ for some coefficients $a_0,..,a_T$, and hence that for all $t = 0,..,T$.

$$\rho(X,X_t) = \frac{E(\Sigma_s a_s X_s X_t)}{\sigma(X_t)\sigma(\Sigma_s a_s X_s)} = \frac{a_t \sigma^2}{\sqrt{\sigma^2}\sqrt{\sigma^2 \Sigma_s a_s^2}} = \frac{a_t}{\sqrt{\Sigma_s a_s^2}} . \tag{4.23}$$

Moreover, $\rho(X,\mathbf{S}) \geq \rho(X^*,\mathbf{S}) > 0$ implies that $a_t > 0$ for all t. Thus we may conclude from Lemma 4.3 that

$$\rho(X,S) \geq \rho(X^*,S) \Rightarrow \min_t \frac{a_t}{\sqrt{\Sigma_s a_s^2}} \geq \frac{1}{\sqrt{T+1}} \tag{4.24}$$
$$\Rightarrow (a_0 = \cdots = a_T) \Rightarrow X = aX^*$$

with $a = a_0\sqrt{T+1} > 0$. Finally, since X is also a $(0,\sigma^2)$-random variable, then $\sigma^2 = E(X^2) = a^2\sigma^2(X^*) = a^2\sigma^2 \Rightarrow a = 1$, so that X^* is the unique $(0,\sigma^2)$-random variable satisfying (4.20). End of proof.

Employing this result, it is now a simple matter to establish (4.14). However, to guarantee uniqueness of this maximum, it is convenient to introduce the following normalization. Let $\Sigma = K(0)$ denote the *autocovariance matrix* for each random vector in the stationary segment $\mathbf{X}_T = [X_0,..,X_T]$, and let each process $\mathbf{Z} = \{Z_t \mid t \in \mathbf{T}\}$ in $\mathbf{P}$ with $\text{cov}(Z_t) = \Sigma$ for all $t \in \mathbf{T}$ be designated as Σ-*process*. Since the original process $\mathbf{X}$ from which $\mathbf{X}_T$ is drawn is necessarily a Σ-process by definition, it is appropriate to require that our finitely-stationary model of $\mathbf{X}$ also be a Σ-process. Hence we now restrict our attention to such processes, and begin by observing from (4.9) that the circular smoothing $\mathbf{C}$ of $\mathbf{X}_T$ is itself a Σ-process. Moreover, from among all such processes in $\mathbf{P}_T^+$, we now have the following characterization of $\mathbf{C}$:

Theorem 4.1 *For any stationary segment* $\mathbf{X}_T = [X_0,..,X_T]$ *from a* Σ*-process in* $\mathbf{P}$, *the circular smoothing* $\mathbf{C}$ *of* $\mathbf{X}_T$ *is the unique best uniform representative of* $\mathbf{X}_T$ *from among all* Σ*-processes in* $\mathbf{P}_T^+$.

Proof: First recall from (2.3) that for all $\mathbf{Z} = \{Z_T \mid t \in \mathbf{T}\} \in \mathbf{P}_T^+$,

$$\rho(\mathbf{X}_T,\mathbf{Z}) = \min\{\rho(X^s_{\tau j}, Z_{\langle s+\tau\rangle j}) \mid \tau = 0,..,T;\ |s| \leq T;\ j = 1,..,n\}\ . \tag{4.25}$$

Moreover, if we now set $t = \langle s+\tau\rangle$ and observe from (4.2) through (4.4) that $0 = \langle s+\tau\rangle - t = \langle \tau+s-t\rangle = \langle s-\langle t-\tau\rangle\rangle \Rightarrow s = \langle t-\tau\rangle$, it follows that the indices in (4.25) may be relabeled as

$$\rho(\mathbf{X}_T,\mathbf{Z}) = \min\{\rho(X^{\langle t-\tau\rangle}_{\tau j}, Z_{tj}) \mid \tau = 0,..,T;\ |t| \leq T;\ j = 1,..,n\} \tag{4.26}$$
$$= \min_{(t,j)} [\min_\tau \rho(X^{\langle t-\tau\rangle}_{\tau j}, Z_{tj})]\ .$$

Hence, observing that for each (t,j) the finite set $\{X^{\langle t-\tau\rangle}_{\tau j} \mid \tau = 0,..,T\}$ consists of uncorrelated $(0,\sigma_j^2)$-random variables (where σ_j^2 is the j-th diagonal element of Σ), it follows at once from Lemma 4.4 and (4.1) that for all Σ-processes $\mathbf{Z} = \{Z_t \mid t \in \mathbf{T}\} \in \mathbf{P}_T^+$, (which in particular satisfy $E(Z_{tj}^2) = \sigma_j^2$ for all t and j) we must have

$$\min_\tau \rho(X^{\langle t-\tau\rangle}_{\tau j}, C_{tj}) \geq \min_\tau \rho(X^{\langle t-\tau\rangle}_{\tau j}, Z_{tj}) \tag{4.27}$$

with equality holding only if $C_{tj} = Z_{tj}$ for all t and j. Thus, for all Σ-processes, $\mathbf{Z} \in \mathbf{P}_T^+$ with $\mathbf{Z} \neq \mathbf{C}$, we may conclude from (4.27), together with the finiteness of the set of indices (t,j), that

$$\min_{\tau} \rho(X_{\tau j}^{\langle t-\tau \rangle}, C_{tj}) > \min_{\tau} \rho(X_{\tau j}^{\langle t-\tau \rangle}, Z_{tj}) \;\; ; \;\; |t| \leq T \;\; ; \;\; j = 1,..,n$$

$$\Rightarrow \min_{(t,j)} \left[\min_{\tau} \rho(X_{\tau j}^{\langle t-\tau \rangle}, C_{tj})\right] > \min_{(t,j)} \left[\min_{\tau} \rho(X_{\tau j}^{\langle t-\tau \rangle}, Z_{tj})\right]$$

$$\Rightarrow \rho(\mathbf{X}_T, \mathbf{C}) > \rho(\mathbf{X}_T, \mathbf{Z})$$

and the result is established. End of proof.

4.3 Relation between Spectra and Periodograms

For any T-stationary process $\mathbf{Z} = \{Z_T \mid t \in \mathbf{T}\}$ in $\mathbf{P}$ with complex spectral representation given by (3.22), each scaled covariance matrix [recall (3.20)]

$$\Gamma_{\mathbf{Z}}(\lambda_\tau) = \frac{1}{2\pi} H_\tau = \left(\frac{2T+1}{2\pi}\right) E(S_\tau S_\tau^*) \;\; ; \;\; \tau \in [T] \tag{4.29}$$

is now designated as the *spectral matrix* for $\mathbf{Z}$ at angular frequency λ_τ, and the full set of such matrices $\{\Gamma_{\mathbf{Z}}(\lambda_\tau) \mid \tau \in [T]\}$ is designated as the $\mathbf{Z}$-*spectrum*. Each spectral matrix $\Gamma_{\mathbf{Z}}(\lambda_\tau)$ summarizes the covariation among the λ_τ-frequency components of the individual processes $\mathbf{Z}_j$ in $\mathbf{Z}$. This covariation may be expressed directly in terms of the covariance kernel K for $\mathbf{Z}$ by simply substituting the definition of H_τ in (3.14) into (4.29). However, a more instructive derivation is made possible by observing from (3.24) that each S_t^* can be expressed in terms of the Z_t's as

$$S_t^* = \frac{1}{2T+1} \sum_{t \in [T]} r_\tau^{t} Z_t' \tag{4.30}$$

Next, post-multiplying (3.22) by S_t^* and recalling that $E(S_\tau S_t^*) = 0$ for all $\tau \neq t$, it follows that $E(Z_t, S_t^*) = r_t^t E(S_t S_t^*)$, and hence that for all $\tau \in [T]$,

$$\begin{aligned} E(S_\tau S_t^*) &= r_\tau^{-\tau} E(Z_\tau S_\tau^*) = r_\tau^{-\tau} \left[\frac{1}{2T+1} \sum_{t \in [T]} r_t^t E(Z_\tau Z_t')\right] \\ &= \frac{1}{2T+1} \sum_t r_\tau^{t-\tau} K(\tau - t) = \frac{1}{2T+1} \sum_t r_\tau^{\langle \tau - t \rangle} K(\tau - t) \\ &= \frac{1}{2T+1} \sum_{h \in [T]} r_\tau^{-h} K(h) \end{aligned} \tag{4.31}$$

where $h = \langle \tau - t \rangle$. Thus recalling (3.4) and (4.9), it follows in particular that for the *circular smoothing* $\mathbf{C}$ of a stationary segment $\mathbf{X}_T$, the associated $\mathbf{C}$-*spectrum* is given by

$$\Gamma_C(\lambda_\tau) = \frac{1}{2\pi} \sum_{h\in[T]} \left(\frac{T+1 - |h|}{T+1}\right) K(h) \exp(-i\,\lambda_\tau\, h) \quad ; \quad \tau \in [T]\,. \tag{4.32}$$

On the other hand, if we consider the *finite Fourier transform*

$$F_T(\lambda) = \sum_{t=0}^{T} X_t \exp(-i\,\lambda\, t) \quad ; \quad -\pi \le \lambda \le \pi \tag{4.33}$$

of a stationary segment $\mathbf{X}_T = [X_0,..,X_T]$ with covariance kernel $\{K(h) \mid h\in[T]\}$ then, following Brillinger (1975; sects. 5.2 and 7.3), we may define the $\mathbf{X}_T$-*periodogram* for all angular frequencies $\lambda_\tau = 2\pi\,\tau/2T+1$ by

$$\begin{aligned} P_T(\lambda_\tau) &= \frac{1}{2\pi(T+1)} E[F_T(\lambda_\tau) F_T(\lambda_\tau)^*] \\ &= \frac{1}{2\pi(T+1)} \sum_{t=0}^{T} \sum_{s=0}^{T} K(t-s) \exp[-i\,\lambda_\tau\,(t-s)] \\ &= \frac{1}{2\pi} \sum_{h\in[T]} \left(\frac{T+1 - |h|}{T+1}\right) K(h) \exp(-i\,\lambda_\tau\, h)\,. \end{aligned} \tag{4.34}$$

Finally, comparing (4.32) and (4.34), we may conclude that

$$\Gamma_C(\lambda_\tau) = P_T(\lambda_\tau) \quad ; \quad \tau \in [T] \tag{4.35}$$

holds identically for all angular frequencies, and hence that: *the* C*-spectrum is precisely the* $\mathbf{X}_T$*-periodogram.* [12]

Moreover, if we recall from (3.31) and (3.32) that for all $\tau = 1,..,T$,

$$\begin{aligned} \Gamma_C(\lambda_\tau) &= \left(\frac{2T+1}{2\pi}\right) E(S_\tau S_\tau^*) = \left(\frac{2T+1}{2\pi}\right) E[(X_\tau + i\,Y_\tau)(X_\tau' - i\,Y_\tau')] \\ &= \left(\frac{2T+1}{2\pi}\right)\left(\frac{1}{4}\right) E[(W_\tau - i\,V_\tau)(W_\tau' + i\,V_\tau')] \\ &= \left(\frac{2T+1}{8\pi}\right)[E(W_\tau\,W_\tau' + V_\tau\,V_\tau') + i\,E[W_\tau\,V_\tau' - V_\tau\,W_\tau')] \end{aligned} \tag{4.36}$$

[12]This identity is made possible by evaluating the $\mathbf{X}_T$-periodogram at the angular frequencies λ_τ of the C-spectrum. Observe, however, from (4.33) that this periodogram is well defined for any frequency in the range $-\pi$ to π. Moreover, since the sample $\mathbf{X}_T$-periodogram is typically evaluated at the *sample-harmonic frequencies* $\theta_\tau = 2\pi\tau/T+1$, $\tau = 1,..,m_T$ in this range (where $m_T = T/2$ for T even and $m_T = (T+1)/2$ for T odd, as in Fuller (1976; sect. 7.1)), it is also of interest to interpret the sample $\mathbf{X}_T$-periodogram at these frequencies. To do so, observe simply that for all $\tau = 1,..,m_T$, $\theta_\tau = 2\pi\tau/T+1 \approx 2\pi(2\tau)/2T+1 = \lambda_{2\tau}$. Hence, the sample $\mathbf{X}_T$-periodogram evaluated at the sample-harmonic frequencies θ_τ may be interpreted as approximate estimates of the C-spectrum at the corresponding angular frequencies $\lambda_{2\tau}$, $\tau = 1,..,m_T$.

then it follows from the identities in (3.27) and (3.28) that the real and imaginary parts of $\Gamma_C(\lambda_\tau)$ are given respectively by:

$$\Gamma_C^c(\omega_\tau) = \frac{2T+1}{8\pi} E(W_\tau W_\tau' + V_\tau V_\tau') \tag{4.37}$$

$$= \frac{2T+1}{4\pi} \text{cov}(W_\tau, W_\tau) = \frac{2T+1}{4\pi} \text{cov}(V_\tau, V_\tau)$$

and,

$$\Gamma_C^q(\omega_\tau) = \left(\frac{2T+1}{8\pi}\right) E(W_\tau V_\tau' - V_\tau W_\tau') \tag{4.38}$$

$$= \frac{2T+1}{4\pi} \text{cov}(W_\tau, V_\tau) = -\frac{2T+1}{4\pi} \text{cov}(V_\tau, W_\tau)$$

where $\omega_\tau = \lambda_\tau/2\pi$ is the time frequency for λ_τ. The sets of matrices $\{\Gamma_C^c(\omega_\tau) \mid \tau = 1,..,T\}$ and $\{\Gamma_C^q(\omega_\tau)) \mid \tau = 1,..,T\}$ are designated, respectively, as the C-*cospectrum* and C-*quadrature spectrum*. Finally, observing that $\exp(-i\,2\pi\omega h) = \cos(2\pi\omega h) - i\sin(2\pi\omega h)$, it follows that (4.34) can be written in terms of (2.10) and (2.11) for each time frequency ω_τ as

$$P_T(\lambda_\tau) = P_T^c(\omega_\tau) + i\,P_T^q(\omega_\tau) \quad ; \quad \tau = 1,..,T\,. \tag{4.39}$$

Hence the real form of the complex identity in (4.35) is easily seen from (4.37) through (4.39) to be given by (2.12) and (2.13).

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation under grant no. SES 87-09312.

REFERENCES

Andersson, T.W., 1971, *The Statistical Analysis of Time Series*, Wiley, New York.
Bachman, G. and L. Narici, 1966, *Functional Analysis*, Academic Press, New York.
Bennett, R.J., 1979, *Spatial Time Series*, Pion Press, London.
Brillinger, D.R., 1975, *Time Series: Data Analysis and Theory*, Holt, Rinehart and Winston, New York.
Cramer, H. and M.R. Leadbetter, 1967, *Stationary and Related Stochastic Processes*, Wiley, New York.
Davis, P.J., 1979, *Circulant Matrices*, Wiley, New York.
Dhrymes, P.J., 1970, *Econometrics: Statistical Foundations and Applications*, Harper and Row, New York.

Duncan, D.B. and R.H. Jones, 1966, "Multiple Regression with Stationary Errors", *Journal of the American Statistical Association*, 61:917-928.

Engle, R.F., 1974, "Band Spectrum Regression", *International Economic Review*, 15:1-11.

Fuller, W,.A., 1976, *Introduction to Statistical Time Series*, Wiley, New York.

Halmos, P.R., *Measure Theory*, Van Nostrand Reinhold, New York,

Hannan, E.J., 1970, *Multiple Time Series*, Wiley, New York.

Harvey, A.C., 1978, "Linear Regression in the Frequency Domain", *International Economic Review*, 19:507-512.

Jenkins, G.M. and D.G. Watts, 1968, *Spectral Analysis and its Application*, Holden-Day, San Francisco.

Smith, T.E., 1981, "Exploratory Analysis of Spatio-Temporal Data Series: with Applications to Regional Unemployment", in P. Friedrich and W. Buhr, (eds.), *Regional Development under Stagnation*, Nomos-Verlag, Baden-Baden, Germany, pp. 319-373.

Streitberg, B., 1979, "Multivariate Models of Dependent Spatial Data", in C.P. Bartels and R.H. Ketellapper, (eds.), *Exploratory and Explanatory Statistical Analysis of Spatial Data*, Martinus Nijhoff, Boston.

Wahba, G., 1968, "On the Distribution of Some Statistics Useful in the Analysis of Jointly Stationary Time Series", *Annals of Mathematical Statistics*, vol. 39, pp. 1849-1862.

Advances in Spatial Theory and Dynamics
A.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 11

Spatial Interaction Models and Their Micro-Foundation

G. Haag

1. INTRODUCTION

During the last decade many efforts have been made to introduce dynamic models into urban regional analysis.

Allen and Sanglier (1978, 1979) developed a spatial interaction model in which jobs and populations of different kinds follow a logistic evolution path in each region and compete with each other via corresponding "capacities". The model includes concepts like "capacities", distance-effects and some economic aspects. Of course, such a logistic model structure could be derived from a stochastic background theory assuming a birth/death structure in the transition rates. But not all processes which are important in spatial dynamics can be fitted into this restrictive frame, since migration or flow-processes are important as well.

The spatial interaction model of Harris and Wilson (1978) is also of a logistic type. The imbalance between revenue attracted to a region and the cost of supply to maintain a certain level of the facility stock leads to changes in the level of the facility stock. Contrary to the above-mentioned model, however, Harris and Wilson take into account expenditure flows, but on the other hand they neglect some important economic factors which are included in the first model. The expenditure flows are directly linked to the facility stock which takes into account an interaction which decreases with distance.

Due to these two important papers geographers and economists became increasingly interested in applying these models to practical situations and comparing the empirical trajectories with the different model outputs. However, three main problems became obvious:

(i) difficulties in calibration,
(ii) the restriction to supply-side orientation,
(iii) the lack of a microeconomic foundation.

(i) The first problem is caused by the nonlinear structure of both models (see Pumain et al., 1986; Lombardo and Rabino, 1983). Considerable difficulties in calibration exist, since the number of trajectories which can be compared with empirical data is less than the number of fitting parameters. Therefore, in a simple application of the models there is no general procedure to guarantee a unique choice of the parameters which have to be estimated. Unfortunately, uncertainty in the model parameters may yield unacceptable differences in the trajectories - even in a short-term simulation.

(ii) Both models focus primarily on the supply side. There are no dynamic equations for the demand-side, which could describe reactions of the expenditure flow to changes in the facility stock.

(iii) The dynamic equations of motion of the models discussed so far are not derived from micro-economic foundations - they originate from the decision behaviour of the different decision makers.

In this paper, we attempt to improve the construction of a spatial-interaction model by using a master equation approach. We generally follow the line of argument in Haag and Wilson (1986). We begin with the decision processes of those individuals who decide to use the facilities of different regions, as well as those of the developers, retailers and land owners who decide about changes in the facility stock, the prices of goods and services and the land rent. We shall not use any equilibrium assumption to derive the fundamental dynamic equations of motion for the relevant macrovariables of the system.

The different decision processes are modelled using a master equation (see Weidlich and Haag, 1983). Fluctuations (uncertainties in the decision process) are introduced via the master equation approach. As a result, we obtain the time-dependent probability that a certain decision configuration (socio-configuration) is realized. The framework also provides the link between micro-economic considerations (socio-economic aspects taken into account in the decision process) and the macro-economic equations of motion for aggregate mean values. These mean values are obtained from the master equation by aggregation with the probability distribution as a weighting factor. This broader overall framework, with full probabilistic dynamic underpinning, also facilitates extensions in a number of directions. For instance, it offers a simple way of introducing saturation effects. The way in which conventional spatial interaction models can be generalized becomes obvious, especially the means of introducing fundamental economic aspects. The master equation offers a new approach to parameter estimation (Haag and Weidlich, 1983; Dendrinos and Haag, 1984; Haag and Dendrinos, 1983); and the known bifurcation properties of spatial interaction models can be used to see how these properties manifest themselves in master equations.

In Section 2, we describe a service system using a notation which is suitable for the master equation approach. In Section 3, the model is formulated. Some concluding comments are offered in Section 4.

2. A SERVICE SYSTEM AS THE BASIS OF THE MODEL

We consider four kinds of "agents" (see Table 1), namely; consumers in region i spending their expenditures $T_{ij}(t)$ in region j, developers controlling facility size $Z_j(t)$ in j, retailers determining the prices $p_j(t)$ of goods or services in region j, and land owners deciding about the land rent $r_j(t)$ in region j. For the purpose of this paper, we shall assume residential location to be fixed, though this assumption can easily be relaxed (as in Haag, 1985).

We shall consider an urban system consisting of L zones with a given transportation network. The transportation costs between zone i and zone j are denoted by c_{ij}. The expenditure configuration is described by the array $\{T_{ij}(t)\} = \underline{T}(t)$ at time t. We will assume that this array is determined by utility functions $v_{ij}(t)$ through a mechanism to be described later. The $T_{ij}(t)$ describe expenditure flows from zone i (consumers' residences) to zone j (where the consumer is using facilities) at time t. Moreover, the utility functions

Table 1: The different 'agents' of the model

'agent'	decision concerning	variable to be controlled via the decision process
consumers	destination choice	$T_{ij}(t)$
developers	facility size	$Z_j(t)$
retailers	price of goods or services	$p_j(t)$
land owners	price of one unit of land	$r_j(t)$

$v_{ij}(t)$ have two subscripts since they relate to both housing and service sectors. The utility functions, $v_{ij}(t)$ are assumed to depend on the supply side variables. Let $Z_j(t)$ be the scale of provision of facilities in zone j at time t and let $p_j(t)$ be the unit price of the service (assuming one type of good or service for simplicity). Let $r_j(t)$ be the land rent to be paid per unit of facilities provided at j.

The modelling task therefore is to determine the expenditure flows $\{T_{ij}(t)\}$ at time t, the spatial and temporal pattern of provision, $\{Z_j(t)\}$, prices $\{p_j(t)\}$ and rents $\{r_j(t)\}$ for given initial conditions $\{T_{ij}(0)$, $Z_j(0)$, $p_j(0)$ and $r_j(0)\}$.

2.1 The Master Equation

Decision processes are stochastic processes, since we are interested in the probability $P(\underline{c};t)$ that a configuration $\underline{c} = P\{\underline{T},\underline{Z},\underline{p},\underline{r}\}$ is realized at time t. The master equation is the equation of motion for this probability distribution.

$$\frac{dP(\underline{c};t)}{dt} = \sum_{\underline{k}} w(\underline{c};\underline{c}+\underline{k})P(\underline{c}+\underline{k};t) - \sum_{\underline{k}} w(\underline{c}+\underline{k};\underline{c})P(\underline{c};t) \qquad (1)$$

with

$$\sum_{\underline{c}} P(\underline{c};t) = 1. \qquad (2)$$

The transition rate from one state $(\underline{c}+\underline{k})$ to another state $(\underline{c})$ is described by $w(\underline{c};\underline{c}+\underline{k})$. The modelling of the transition rate $w(\underline{c};\underline{c}+\underline{k})$ in terms of socio-economic processes is the main task of the application of the master equation framework to socio-economic processes. This will be done in the next section. Let us now consider how to obtain the macro-economic equations of motion if the micro-economic decision processes are specified.

2.2 The Macro-economic Equation of Motion

Equations of motion for the means values (denoted by a bar) of the expenditure flows $\bar{T}_{ij}(t)$, the facility stock $\bar{Z}_j(t)$, the prices $\bar{p}_j(t)$ and land rents $\bar{r}_j(t)$ can be obtained from the master equation by multiplying (1) with T_{ij}, Z_j, p_j, r_j, respectively and by summation over all possible configurations $\underline{c} = (\underline{T}, \underline{Z}, \underline{p}, \underline{r})$ (cf. Weidlich and Haag, 1983).

2.3 The Total Transition Rate

The total transition rate $w(\underline{c}+\underline{k}; \underline{c})$ is obtained as the sum over contributions of different socio-economic processes.

$$w(\underline{c}+\underline{k}; \underline{c}) = \sum_{i,j,k}^{L} w_{ik,ij}(\underline{T}+\underline{k}_1, \underline{Z}, \underline{p}, \underline{r}; \underline{T}, \underline{Z}, \underline{p}, \underline{r}) + \sum_{j,\ell}^{L} [w_j^{+(\ell)}(\underline{c}+\underline{k}; \underline{c}) + w_j^{-(\ell)}(\underline{c}+\underline{k}; \underline{c})] \quad \text{for} \quad \ell = 2, 3, 4 \tag{3}$$

The $w_{ik,ij}(\underline{T}+\underline{k}_1, \underline{Z}, \underline{p}, \underline{r}; \underline{T}, \underline{Z}, \underline{p}, \underline{r})$ refer to changes in the expenditure flow configuration $\underline{T}$, due to decisions by consumers to change from a state which implies residence in i and using facilities in j, to a state with residence in i but now using facilities in k. Thus changes of residential location are not taken into account, since it is reasonable to assume that housing mobility will be considerably slower than the shopping mobility of the population. Indeed, decisions to buy in another zone can be taken much more easily than those to change an apartment (Leonardi, 1985).

The $w_j^{+(\ell)}, w_j^{-(\ell)}$ describe decisions of entrepreneurs ($\ell = 2$), retailers ($\ell = 3$) and land owners ($\ell = 4$) to add or to remove one unit of the facility stock Z_j, the prices p_j and the rents r_j, respectively.

The crucial task is now to specify these transition rates. Our socio-economic experience must be used to model these decision processes.

2.4 Decisions of consumers

The actual expenditure flow array

$$\underline{T}(t) = \{T_{11}(t), T_{12}(t), \ldots, T_{1L}(t), T_{21}(t), \ldots T_{LL}(t)\} \tag{4}$$

will change in the course of time due to decisions by consumers to use facilities in another zone.

The total revenue, $D_j(t)$, attracted to each zone j is obtained by

$$D_j(t) = \sum_{i=1}^{L} T_{ij}(t)\,; \tag{5}$$

the total expenditure $E_i(t)$ of consumers living in zone i is

$$E_i(t) = \sum_{j=1}^{L} T_{ij}(t). \tag{6}$$

The transition rate $p_{ik,ij}(\underline{T},\underline{Z},\underline{p},\underline{r})$ of an individual living in i, for changing their shopping area $j \rightarrow k$, is now assumed to depend on the expected utility gain $(v_{ik}-v_{ij})$

$$p_{ik,ij}(\underline{T},\underline{Z},\underline{p},\underline{r}) = \varepsilon_1^*(t)\exp[v_{ik}(t) - v_{ij}(t)] > 0 \tag{7}$$

where $\varepsilon_1^*(t)$ is a time scaling parameter.

Let $n_{ij}(t)$ be the number of individuals having residence in zone i, i=1,2,...,L, and using facilities in j. Then

$$n_i(t) = \sum_{j=1}^{L} n_{ij}(t) \tag{8}$$

is the number of consumers living in i, and the total number of all consumers in the city areas is

$$N(t) = \sum_{i=1}^{L} n_i(t). \tag{9}$$

Then we obtain for the total transition rate per unit of time (changes in the shopping trips of the population per time interval)

$$\varpi_{ik,ij}(\underline{c}+\underline{k};\underline{c}) = n_{ij}(t)\, p_{ik,ij}(\underline{c}). \tag{10}$$

In many applications it may be reasonable to assume that the expenditure flows $T_{ij}(t)$ are related to the consumers' activities $n_{ij}(t)$ by

$$T_{ij}(t) = g(t)n_{ij}(t) \tag{11}$$

where g(t) can be seen as the average of individual needs. By comparing the total stocks, g(t) can be easily determined.

Thus the contribution of the consumers to the total transition rate per unit of time (changes in the expenditure flows) finally reads

$$w_{ik,ij}(\underline{T}+\underline{k}_1,\underline{Z},\underline{p},\underline{r};\underline{T},\underline{Z},\underline{p},\underline{r}) = \begin{cases} T_{ij}\varepsilon_1(t)\exp[v_{ik}(t) - v_{ij}(t)] \\ \quad \text{for } \underline{k}_1 = \{0,...,1_{ik},..0,...-1_{ij}..\}. \\ 0 \text{ for all other } \underline{k}_1. \end{cases} \tag{12}$$

2.5 The Decision Processes of Developers, Retailers and Land Owners

The dynamics of the provision of facilities, $Z_j(t)$, prices, $p_j(t)$, and rents, $r_j(t)$, can be considered as an analogy to birth-death processes.

Let $\underline{X}^{\ell} = (X_1^{\ell}, X_2^{\ell},..., X_L^{\ell})$ be the facility stock configuration ($\ell = 2$; $X_j^{(2)} = Z_j$), the price configuration ($\ell=3$; $X_j^{(3)} = p_j$), or the rent configuration ($\ell= 4$; $X_j^{(4)} = r_j$), which

respectively describe one possible realization of a socio-economic distribution of facilities, prices or rents, at time t. Due to changing decisions by the corresponding decision maker, the configuration $X_j^{(\ell)}$ will change over the course of time. We shall now introduce the rates corresponding to these processes in the same manner as above.

Let $w_j^{+(\ell)}(\underline{X}+\underline{k}^{(\ell)}; \underline{X})$, $w_j^{-(\ell)}(\underline{X}+\underline{k}^{(\ell)}; \underline{X})$ be the 'growth' rate, and 'death' rate per unit of time of the stock variable $\underline{X}^{(\ell)}$. Then a rather general formulation for these rates may be the following (cf. Haag and Wilson, 1986)

$$w_j^{+(\ell)} = \frac{1}{2} f_j^{(\ell)}(\underline{X}) \exp \phi_j^{(\ell)}(\underline{Z},\underline{p},\underline{r}) > 0 \tag{13a}$$

$$w_j^{-(\ell)} = \frac{1}{2} f_j^{(\ell)}(\underline{X}) \exp [-\phi_j^{(\ell)}(\underline{Z},\underline{p},\underline{r})] > 0, \tag{13b}$$

for $\ell = 2, 3, 4$ and for rather general functions $f_j^{(\ell)}(\underline{X}) > 0$ and $\phi_j^{(\ell)}(X)$. Of course, an immigration term could be included in the transition rates $w_j^{+(\ell)}(\underline{X}+\underline{k}^{(\ell)}; \underline{X})$ to describe the possible settlement of facilities in initially empty zones. The functions can be modelled to include saturation effects of the various parameters. The factor $f_j^{(\ell)}(X)$ describes the speed of adjustment. Since the birth-death rates must not be negative, the condition $f_j^{(\ell)}(X) > 0$ has to be fulfilled. Here we assume for simplicity that the speed of adjustment depends on the scale of the stock $\underline{X}^{(\ell)}$ in a linear way, with time scaling parameters $\varepsilon_\ell(t)$ such that

$$f_j^{(\ell)}(\underline{X}) = \varepsilon_\ell(t)\, X_j^{(\ell)} . \tag{14}$$

The function $\phi_j^{(\ell)}(X)$ takes into account the imbalance between the cost of supply and the revenue attracted to facilities in the zone under consideration. If there is an economic surplus, $\phi_j^{(\ell)}(X) > 0$, it is more likely that the facility stock will expand, the prices and the land rents will increase; and vice versa. Reasonable assumptions are:

$$\phi_j^{(\ell)} = \lambda_\ell [D_j(t) - C_j(t)] \tag{15a}$$

or alternatively

$$\phi_j^{(\ell)} = \lambda_\ell [D_j(t) - C_j(t)]/Z_j(t), \tag{15b}$$

where λ_ℓ describes the intensity of response of an 'agent' to an economic surplus. The total revenue attracted to j is D_j. The cost of supplying facilities of size $Z_j(t)$ at zone j, when land rent is $r_j(t)$, is denoted by $C_j(t)$. We assume that if there is an economic surplus $D_j(t) > C_j(t)$, in j, the probability of an extension of the facility stock exceeds the probability of a reduction due to the experience of the decision-makers.

The difference between (15a) and (15b) is that in (15a) the decision function $\phi_j^{(\ell)}(X)$ is assumed to be proportional to the economic surplus in the zone j whereas in (15b) a proportionality to the economic surplus of the facility stock is considered. We assume further that the cost of supply is proportional to the number of units of the facility stock

$$C_j(t) = [k_j + r_j(t)]\, Z_j(t) \tag{16}$$

where k_j is a fixed constant and r_j is a measure for the rent of one unit of facilities in zone j.

3. THE MEAN VALUE EQUATIONS OF THE DYNAMIC SERVICE SECTOR MODEL

By substitution of the transition rates (12), (13), with (3) into the master equation (1), we obtain the evolution over time of the probability distribution $P(\underline{c}; t)$ from a given initial distribution $P(\underline{c}; 0)$. Since it is difficult to handle this problem in practice - even numerically - we derive the more aggregate set of mean value equations for our service sector model from the master equation (1) assuming a unimodal and sharply peaked distribution function:

$$\dot{\bar{T}}_{ij}(t) = \varepsilon_1(t)\ \Big\{ \sum_{k=1}^{L} \bar{T}_{ik}(t) \exp[v_{ij}(t) - v_{ik}(t)] - \sum_{k=1}^{L} \bar{T}_{ij}(t) \exp[v_{ik}(t) - v_{ij}(t)] \Big\}, \tag{17}$$

$$\dot{\bar{Z}}_j(t) = \varepsilon_2(t)\, \bar{Z}_j(t) \sinh\Big[\lambda_2 \Big(\sum_{i=1}^{L} \bar{T}_{ij}(t) - (k_j + \bar{r}_j(t))\, \bar{Z}_j(t)\Big) / \bar{Z}_j^{\delta}\Big], \tag{18}$$

$$\dot{\bar{p}}_j(t) = \varepsilon_3(t)\, \bar{p}_j(t) \sinh\Big[\lambda_3 \Big(\sum_{i=1}^{L} \bar{T}_{ij}(t) - (k_j + \bar{r}_j(t))\, \bar{Z}_j(t)\Big) / \bar{Z}_j^{\delta}\Big], \tag{19}$$

$$\dot{\bar{r}}_j(t) = \varepsilon_3(t)\, \bar{r}_j(t) \sinh\Big[\lambda_4 \Big(\sum_{i=1}^{L} \bar{T}_{ij}(t) - (k_j + \bar{r}_j(t))\, \bar{Z}_j(t)\Big) / \bar{Z}_j^{\delta}\Big], \tag{20}$$

for i, j = 1,2,...L, $\delta = 0$ or 1.

Equations (17)-(20) together with (5), (15) and (16) complete our model (the utility functions v_{ij} will be specified in the next subsection). For given initial conditions ($T_{ij}(0)$, $Z_j(0)$, $p_j(0)$, $r_j(0)$) the trajectories for $t > 0$ can be computed. The hyperbolic sine function leads to an amplification of the reactions of the decision makers on economic disequilibrium. Near equilibrium (18), the Harris and Wilson hypothesis is yielded. However, the estimation of the model parameters from empirical data can now be performed by directly linking the transition rates (12), (13) to the corresponding empirical results, eg.by applying a generalized least square procedure (Haag, 1985; a detailed description of the estimation procedure is given in Weidlich and Haag, 1987). This interesting problem of parameter estimation will not, however, be discussed in this paper.

The dynamic service sector model (17)-(20) exhibits many different time scales. For example, the time scale for shopping trips $\varepsilon_1(t)$, is really quite different from those changing the facility stock $\varepsilon_2(t)$, the price dynamics $\varepsilon_3(t)$, and the rental dynamics $\varepsilon_4(t)$. Customers have no long-term commitments to the shops, and can easily change their choices every day. This very fast process can therefore be reasonably described by its steady state. This may lead to analytical simplifications of the model and is called adiabatic elimination.

Using this adiabatic elimination procedure, the $T_{ij}(t)$ are driven by the dynamics of $Z_j(t)$, $r_j(t)$ and $p_j(t)$.

3.1 The Stationary Solution of the Service Sector Model

Considering the stationary version of (17), we find for the stationary expenditure flows, which we denote by $\hat{T}_{ij}$, that:

$$\hat{T}_{ij} = [\hat{E}_i \exp(2\hat{v}_{ij})] / [\sum_{k=1}^{L} \exp(2\hat{v}_{ik})]. \tag{21}$$

Equation (21) can easily be proved by inspection. By comparing (21) with the special form of the flow pattern assumed by Harris and Wilson (1985), the corresponding utility functions can be derived

$$v_{ij}^{(w)} = \frac{1}{2}\alpha \ln Z_j(t) - \frac{1}{2}\gamma \ln p_j(t) - \frac{1}{2}\beta\ c_{ij}(t). \tag{22}$$

Near equilibrium $|\phi_i^{(\ell)}(X)| << 1$, with (15a) and assuming that the expenditure flows have reached their equilibrium pattern (21), and that the utility functions (22) are adequate, the Harris and Wilson model is obtained (prices and rents are not considered).

$$\bar{T}_{ij}(t) = (\bar{E}_i(t)\, \bar{Z}_j^{\alpha}(t) \exp[-\beta c_{ij}(t)])/(\sum_{k=1}^{L} \bar{Z}_k^{\alpha}(t) \exp[-\beta c_{ik}(t)]) \tag{23a}$$

$$\dot{\bar{Z}}_j(t) = \epsilon_2^*(t)\bar{Z}_j(t)\,[\sum_{i=1}^{L} \bar{T}_{ij}(t) - k_j\, \bar{Z}_j(t)]. \tag{23b}$$

From an economic point of view, however, the prices of goods p_j and transportation costs c_{ij} should not be separated as they are defined in the utility function (22). Therefore, another reasonable assumption (Haag, 1985) is proposed for the utility function.

$$v_{ij}^{(H)}(t) = \frac{1}{2}\alpha\, Z_j(t)\,[1-Z_j(t) / Z_{j\,sat}] - \frac{1}{2}\gamma\, p_j(t) - \frac{1}{2}\beta\ c_{ij}(t). \tag{24a}$$

This assumption corresponds to a Taylor expansion of the utility function v_{ij} in the state variables. Such an expansion has proved to be powerful in migration theory (Haag and Weidlich, 1984; Weidlich and Haag, 1987). A possible saturation of a zone j with respect to the facility stock Z_j is also taken into account. The saturation level of the facility stock is called $Z_{j\,sat}$. Since both p_j, c_{ij}, are prices it is reasonable to assume $\gamma = \beta$. Then the utility function reads

$$v_{ij}^{(A)}(t) = \frac{1}{2}\alpha\, Z_j(t)\,[1-Z_j(t) / Z_{j\,sat}] - \frac{1}{2}\beta\,(p_j(t) + c_{ij}(t)). \tag{24b}$$

Therefore, in the utility function $v_{ij}^{(A)}$ of consumers having residence in zone i and using facilities in zone j, the perceived prices of goods $(p_j + c_{ij})$ appear.

Assuming that the shopping attitude of individuals living in any other residential area k should not have an important influence on the shopping decisions of individuals living in zone j ($j \neq k$) and thus could be neglected in a first approximation, Frankhauser (1986) proposed another plausible assumption for the utilities

$$v_{ij}^{(F)}(t) = \frac{1}{2}\ell n\,[k_j+r_j)\,Z_j(t)] - \frac{1}{2}\beta\,c_{ij}(t) - \frac{1}{2}\ell n\;\tau_j(t) \tag{25a}$$

with

$$\tau_j(t) = \sum_{i=1}^{L} \exp\,[-\beta\,c_{ij}(t)]. \tag{25b}$$

By insertion of (25) in (21) we obtain a set of decoupled stationary expenditure flows

$$\hat{T}_{ij}^{(F)} = (k_j+\hat{r}_j)\,\hat{Z}_j\,(\hat{\tau}_j)^{-1}\exp\,[-\beta\,\hat{c}_{ij}]. \tag{26}$$

The effect on an improvement of the infrastructure between two zones due to a reduction in the transportation costs between these zones and/or the effect of a sudden increase of the scale of provision of facilities in one particular zone for different values of the deterrence parameter β are simulated and discussed in Haag and Frankhauser (1987).

The facility stock, the prices and the rents have reached their simultaneous equilibrium values, namely the stationary solution of (17)-(20), if the condition

$$\hat{D}_j = \hat{C}_j \tag{27}$$

is fulfilled. This assumption seems to be reasonable for the facility stock but is is not so obvious for the prices and the rents. Hence, further investigations should deal with these economic effects in particular.

4. CONCLUDING COMMENTS

The application of this model to practical situations and its possible use as a planning instrument will be amongst the next steps in our research program. Here we have stressed the embedding of the conventional spatial interaction model into a stochastic framework. A probabilistic description of the decision processes of 'agents' was introduced. The master equation provided the link between the microeconomic level of the individual decision processes to the aggregated dynamic equations of motion for observable mean values. It became obvious that such spatial interaction models can be extended and, even more importantly, linked to socio-economic data. But much work still has to be done to improve the economic aspects of the price and rent adjustment process.

ACKNOWLEDGEMENT

This work was supported by Volkswagen-Stiftung and Deutsche Forschungsgemeinschaft (SFB 230).

The author is also very grateful to Professors W. Weidlich and Å.E Andersson for many stimulating discussions and a critical reading of the manuscript.

REFERENCES

Allen, P.M. and M. Sanglier, 1978, "Dynamic Models of Urban Growth", *Journal of Social and Biological Structures*, 1: 265-280.

Allen, P.M. and M. Sanglier, 1979, "Dynamic Models of Urban Growth", *Journal of Social and Biological Structures*, 2: 269-278.

Birkin, M. and A.G. Wilson, 1985, "Some Properties of Spatial-structural-economic-dynamic models", Working Paper 440, School of Geography, University of Leeds.

Dendrinos, D. and G. Haag, 1984, "Towards a Stochastic Dynamic Theory of Location: Empirical Evidence", *Geographical Analysis*, 16: 287-300..

Frankhauser, P. 1986, "Entkopplung der stationären Lösung des Haag-Wilson-Modells durch einen neuen Ansatz für die Nutzen-Funktion", Arbeitspapier 2. Institut für Theoretische Physik, Universität Stuttgart.

Haag, G., 1985, "Services 2 - A Master Equation Approach", in C.S. Bertuglia, G. Leonardi and A.G.Wilson, (eds.), *Urban Systems: Designs for an Integrated Dynamic Model*, forthcoming.

Haag, G. and D. Dendrinos, 1983, "Towards a Stochastic Dynamical Theory of Location: A Nonlinear Migration Process", *Geographical Analysis* 15: 269-286.

Haag, G. and P. Frankhauser, 1987, "A Stochastic Model of Intraurban Supply and Demand Structures", in H.J.P Timmermans, (ed.), *Contemporary Developments in Quantitative Geography*, forthcoming.

Haag, G. and W. Weidlich, 1984, "A Stochastic Theory of Interregional Migration", *Geographical Analysis*, 16:331-357.

Haag, G. and A.G. Wilson, 1986, "A Dynamic Service Sector Model - A Master Equations Approach with Prices and Land Rents", Working Paper 447, School of Geography, University of Leeds.

Harris, B. and Wilson, A.G., 1978, "Equilibrium Values and Dynamics of Attractiveness Terms in Production-constrained Spatial Interaction Models", *Environment and Planning A* 10: 371-388.

Leonardi, G., 1985, "Housing 3 - Stochastic Dynamics", in C.S. Bertuglia, G. Leonardi and A.G. Wilson, (eds.), *Urban Systems: Designs for an Integrated Dynamic Model*, forthcoming.

Lombardo, S.R. and G.A. Rabino, 1983, "Nonlinear Dynamic Models for Spatial Interaction: The Results of Some Numerical Experiments", Paper presented at the 23rd European Congress, Regional Science Association, Poitiers.

Pumain, D., Th. Saint-Julien and L. Sanders, 1986, "Dynamics of Spatial Structure in French Urban Agglomerations", to be published in: *Papers of the Regional Science Association.*

Weidlich, W. and G. Haag, 1983, *Concepts and Models of a Quantitative Sociology: The Dynamics of Interacting Populations,* Springer-Series in Synergetics 14, Berlin.

Weidlich, W. and G. Haag, (eds.), 1987, *Interregional Migration - Dynamic Theory and Comparative Evaluation*, Springer Series in Synergetics, forthcoming.

Advances in Spatial Theory and Dynamics
Å.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 12

Modelling Non-linear Processes in Time and Space

W. Barentsen and P. Nijkamp

1. INTRODUCTION

In recent years there has been a growing interest in *non-linear dynamic systems* among economists. Several factors lie behind this 'upswing' in attention.

First, the path-breaking *methodological contributions* to (in)stability and (dis)equilibrium analysis in the natural science field - made, among others, by Thom (1975) and Nicolis and Prigogine (1977) - stimulated a thorough investigation into the nature of non-linear dynamics in the social sciences (see e.g., Weidlich and Haag, 1983). It was increasingly acknowledged that dynamic interactions between the components of a complex system marked by dissipative structures affecting inter alia the homogeneity of time and space may lead to a large spectrum of evolutionary patterns of such a system (ranging from inert and stable behaviour to fluctuating and unstable behaviour).

Also, *structural changes* in the economic conditions of western societies have led to increased interest in non-linear evolutionary patterns. Long-wave patterns in macro-economic and regional systems, long-term drastic shifts in economic activities, and differential dynamic trajectories of various subsystems of the economy have demonstrated the relevance of non-linear approaches in economics. In fact, as soon as parameters of an otherwise linear system are time-dependent with respect to endogenous variables of this system, one faces a situation of endogenously determined structural change leading to non-linear dynamic models. In a space-time context such structural changes may lead to interesting situations related to the spatio-temporal (ir)reversible trajectories (including catastrophic behaviour) of a dynamic system.

Finally, for many years the mathematical-statistical difficulties inherent in non-linear dynamic systems have precluded many researchers from applying such approaches to the social sciences. However, the *rapid computational advances* in this field (including operational computer software) and the *availability of an appropriate mathematical framework* (notably the analysis of the qualitative behaviour of a dynamic system) have led to an increased use of non-linear dynamic models in economics.

Examples of such applications can be found in various fields of economics, such as macro-economics (e.g., long waves analysis), consumer economics (e.g., shopping behaviour), regional economics (e.g., urban life cycle analysis) and business economics (e.g. , technology innovation behaviour). In this respect the relationship between the micro behaviour of the system's components and their macro consequences for the system as a whole is intriguing: changes at a micro level may - beyond a certain critical

threshold level - exert structural influences at an aggregate level. Clearly, linear models are in general unable to generate structural or sometimes discontinuous changes.

The class of non-linear dynamic models has specific features which distinguish them from conventional linear dynamic models. The main characteristic is that such models are able to describe qualitative system changes, as a non-linear dynamic model may contain certain ranges of parameter values for which the system can be in *multiple equilibrium states*. Such ambiguity, which in any case does exist in a formal sense, can only be eliminated if either the theory underlying the non-linear dynamic model specification is made more specific (i.e., more oriented toward the behaviour of the system under these parameter values) and hence more aligned to the phenomena to be studied (to diminish the semantic insufficiency), or if more insight is gained into the long-term historical evolution of the phenomenon at hand (requiring full and non-trivial information about past behaviour). In general, however, economic theories are semantically insufficient to avoid a priori the possibility of multiple equilibrium states. The existence of various types of feedback mechanisms in economic systems may lead to non-linear trajectories and even discontinuous changes. Such discontinuous changes are often time-irreversible; i.e., by reversing the direction of the initial stimulus that has caused the discontinuous movement (e.g., bifurcation, catastrophe, or shock), the system does not necessarily move back to its original state. Such asymmetric behaviour implies an unstable evolutionary pattern, as the discontinuities which may then be triggered by marginal changes in initial conditions or in parameter values make the system's evolution time-irreversible. Consequently, the past state of a system plays a dominant and non-trivial role in non-linear dynamic systems.

It is worth noting that non-linear dynamics plays a crucial role in explaining the spatio-temporal evolution of a *spatial* system (e.g., city, region), as the question of isotropy of space and time is at stake here. The analysis of the development of geographical structures requires an investigation into the existence of reversibility of space-time systems. An abstract representation of a geographical structure can be given by the Cartesian coordinates (x, y) of the successive phenomena to be modelled in order to position them in a two-dimensional surface. Additional dimensions (e.g., z) may of course be added to account for other attributes of such a phenomenon: for instance, its size or magnitude, its degree of spatial interaction with respect to other phenomena in space and time, etc. In a general sense geographical structure refers to the interrelatedness between locational aspects (x_i, y_i) and other dimensions z_i of a phenomenon i. Clearly, such structures are the results of a historical process (e.g., investment and locational decisions). This might inter alia be described by means of event-history analysis (see Hannan and Tuna, 1985) in a discrete sense, or by means of continuous space-time models (see Beckmann and Puu, 1985) in a continuous sense. In all these cases the structure and evolution of geographical systems may be analyzed by means of non-linear dynamic models exhibiting discontinuities and irreversibilities. To illustrate the relevance of such approaches, one may quote Griffith and Lea (1983) who remark: "Geographical systems, such as school systems, and geographical networks such as grain elevator and gas station networks, experiencing rationalization, growth or concentration and decline, have demonstrated empirically the asymmetry of life-cycle trajectories".

Section 2 contains some remarks about dynamic systems. Different forms of bifurcation are treated briefly in Section 3. In Section 4 we distinguish three levels on which dynamic processes can be modelled. These three levels are then used in Section 5 to classify some models which have been developed by economists and geographers. Finally, Section 6 contains some conclusions.

2. DYNAMIC SYSTEMS

Following Samuelson (1948, p. 314) and Frisch (1935-36) we may define a dynamical system as a system whose behaviour over time is determined by functional equations in which variables are involved in an essential way at different points of time. This definition becomes more precise when applied to economics due to the requirement that the variables should be economically significant. Otherwise, every variable can be written as the derivative of its own integral, which in itself may not be a variable of interest, although it would make the system dynamic. In respect to this qualification, we regard systems of difference and differential equations as dynamic systems. In the rest of this paper we shall restrict attention to these two classes of dynamic equations.

There are various ways of classifying dynamic models. For instance, Samuelson subdivided them into complete causally-determined systems, historical systems, and stochastic (historical and non-historical) systems (for details see Samuelson, 1948).

The analytical knowledge of systems of differential equations is better developed than that of systems of difference equations. However, several results for systems of differential equations hold in an analogous way for systems of difference equations. Sometimes, however, unexpected results may take place; viz., if differential equations are discretely approximated by means of difference equations. The problem here is caused by the fact that empirical economic data are usually only available at discrete time intervals, so that in economic research one is forced to use difference equations for dynamic systems. This problem can be clarified as follows. Assume the following dynamic system:

$$\dot{x} = F(x) \tag{2.1}$$

where $x \in X \subset IR^n$, $\dot{x} \equiv \frac{dx(t)}{dt}$ and $F: X \rightarrow IR^n$. X is called the state space and F defines a vector field on X, while (2.1) is a general formulation of a system of differential equations.

In applications where numerical solutions are required, the distinction between differential and difference equations becomes blurred, as the computer solution of (2.1) usually requires a discrete approximation:

$$\frac{x(t+\Delta) - x(t)}{\Delta} = F(x_t), \tag{2.2}$$

where $t = 0,\Delta,2\Delta,...$ with Δ a small positive number. Let $n = 1$; local stability of the corresponding equilibrium x^* of (2.1) requires:

$$\left(\frac{dF}{dx}\right)\Big|_{x=x^*} < 0, \tag{2.3}$$

while local stability of the corresponding fixed point x^* of (2.2) requires:

$$\left| 1 + \Delta \frac{dF}{dx}_{x=x^*} \right| < 1. \tag{2.4}$$

This illustrates that a discrete numerical solution procedure for a differential equation system evokes problems of stability: if Δ is taken too large, the fixed point x^* of the

approximated system (2.2) may exhibit unstable behaviour, although it may be a stable equilibrium point for system (2.1).

A glaring example of such a situation can be found in models of the type developed by May (1974), which are frequently used in population dynamics (see also Pimm, 1982; Li and Yorke, 1982; and Brouwer and Nijkamp, 1985).

The prototype of the May model has the following form:

$$x(t+1) = \psi x(t)[1 - x(t)]. \tag{2.5}$$

This simple non-linear dynamic system in difference equation form may exhibit a remarkable spectrum of dynamic behaviour ranging from stability to fluctuating and even chaotic patterns, depending on the parameter values and on the initial conditions. This unusual and unexpected behaviour of a non-linear dynamic model does not, however, hold for its continuous counterpart in differential equation form. Consequently, one may conclude that the May model mainly derives its unusual results from its specification in difference equation form. Although such models may generate a wide spectrum of dynamic behaviour, there is *a priori* no reason to believe that simple models of this type are able to provide a more realistic and reliable representation of a dynamic complex world than other models.

With regard to models of type (2.1) it is interesting to note 7 fundamental questions raised inter alia by Varian (1981):

(i) Do solutions exist?
(ii) Do equilibria exist?
(iii) What is the number of equilibria?
(iv) Which equilibria are locally stable?
(v) Which equilibria are globally stable?
(vi) Do cycles exist?
(vii) Is the system structurally stable?

We shall treat these questions concisely in order to clarify some relevant aspects of (non-linear) dynamic models.

A *solution* to (2.1) - with initial conditions $x(0) = x_0$ - is a differentiable function $x: I \rightarrow X$, where I is an interval in IR, such that:

$$\frac{dx(t)}{dt} = F(x(t)), \text{ and } x(0) = x_0 . \tag{2.6}$$

If F is continuously differentiable on the open subset X, a unique solution does exist. This solution is continuous (as a function of x_0). However, an explicit analytical solution to (2.1) is usually difficult to find, so attention is normally focussed on the qualitative properties of this system. In this framework, the issue of the existence and the stability of equilibria of the system emerges.

An *equilibrium* is defined as a point $x^* \in X$ such that $F(x^*) = 0$. Various theorems dealing with the existence and *number of equilibria* - especially in case of non-linear dynamic models - may be relevant in concrete situations (see, for instance, Rijk and Vorst, 1982). The notion of *local stability* is particularly important in this context. An equilibrium point x^* is called locally (asymptotically) stable, if there is some $\varepsilon > 0$ such that for all x_0 for which $| x_0 - x^* | < \varepsilon$ it follows that $\lim_{t \rightarrow \infty} \varphi_t(x_0) = x^*$, where $\varphi_t(x_0)$ is the flow of the differential equation (2.1) that corresponds to the initial condition $x(0) = x_0$. Usually, only locally stable equilibrium points are regarded as relevant from an economic point of view. Stable equilibria act as *attractors* of the trajectory of a dynamic system and thus determine a specific solution. The case of *unstable* equilibria is also interesting, as such points may act as *repellors* of the trajectories. Finally, a third type of equilibrium is the *saddle point*, which reflects a special kind of instability: a saddle point implies that

there are two trajectories leading to different equilibrium points; such a point divides the state space into two areas, while each trajectory in each area is directed towards a different stable equilibrium. The nature of an equilibrium point x* can in principle be evaluated by analyzing the eigenvalues of the Hessian matrix DF(x*) (see Annex 1).

An equilibrium point x* is *globally stable* if $\lim_{t\to\infty} x(t) = x^*$ for any initial condition x_0. Further contributions to the analysis of global stability in economics can be found in Arrow and Hahn (1971).

A special type of equilibrium is a *cyclical* pattern: a point set X is in a *cycle* (closed orbit), if $F(x) \neq 0$ and $\varphi_t(x) = x$ for some $t \neq 0$. Casti (1985) indicates why cycles are relevant in modelling real-world phenomena. In his view, empirical evidence indicates that for many phenomena periodicity is the rule and static equilibrium the exception. Besides, he states that a system that can respond more swiftly to the environment than its neighbours has a competitive advantage. That real-world systems do not always exhibit truly periodic behaviour is due to perturbations which continually push the system from one cycle toward another.

It is worth noting that the equations of a model always have an approximate character, so that structural stability is a desirable property; i.e., a small change (perturbation) in $F(\cdot)$ should not change the qualitative nature of the vector field. Structural stability is thus directly related to the 'behaviour' (in terms of location, existence and character) of the equilibrium points. Clearly, under specific circumstances with dramatic or discontinuous changes in a real-world system one may use modelling experiments based on structural instability. In the latter case, linear models with endogenous changes cannot be employed.

3. NON-LINEARITY AND BIFURCATIONS

Linear model specifications have become very popular in economics, although there are *a priori* no strong theoretical arguments in favour of linear models. Clearly, practical reasons may be relevant in this context (for instance, data availability, econometric estimation and test procedures, first-order Taylor approximations, computer software like linear programming etc.) but, as noted above, linear models are often inadequate when used in relation to various real-world phenomena.

Non-linear models are able to generate non-trivial changes in a dynamic system (not just growth or decline), and consequently the (event) history of phenomena plays a crucial role in dynamic modelling efforts for such a system. The capability of non-linear models to describe and/or to endogenously generate bifurcations is their major discriminating feature. For instance, by means of bifurcation analysis one may try to model structural change processes: the qualitative nature of the change caused by a bifurcation reveals the evolution of the system concerned.

Consider system (2.1) in a slightly modified form:

$$\dot{x} = F(x;\alpha), \tag{3.1}$$

where α is a parameter vector. Let F be non-linear. Assume now that $\bar{x}_i$ (x=1,...,I) are I distinct equilibrium points. The number I may then be co-determined by the numerical value of α. Since $\dot{x} = 0$ for any point $\bar{x}_i$, there exists an internal consistency within the system for these values of x. A state is *internally consistent*, if it is self-sustaining. The nonlinearity of F may thus lead to a number of distinct self-sustaining states. Existence of cycles reveals that there may also be an internal consistency between moving variables. If one takes for granted that a social system can be decomposed into individuals whose behaviour influences the system's environment and is in turn influenced by this environment, it is easily seen that multiple self-sustaining states may exist.

Each individual action or each local intervention in a complex system may lead to an aggregate impact at the system level that - after a bifurcation - may result in global

changes. Various forms of bifurcation may be distinguished, some of which will be discussed here. Assume for instance the following partition of (2.1)

$$\begin{aligned} &(a)\ \dot{x}_1 = f_1(x_1, x_2; \beta) \\ &(b)\ \dot{x}_2 = f_2(x_2; \gamma) \end{aligned} \tag{3.2}$$

where x_1 and x_2 are vectors of fast respectively slow moving varibles, and β and γ parameter vectors. The functions f_1 and f_2 are assumed to be non-linear.

Since f_1 is non-linear, there may be some ranges of x_2 and β for which (3.2.a) has multiple equilibria, and there may be some bifurcation points (x_2^b, β^b) at which this *number* of equilibria may even change. If the system is situated near such a point, the equilibrium point which will be reached depends on infinitesimal changes in x_2 and/or β (note that x_1 is assumed to be a fast moving variable). Consequently, at such points (x_2^b, β^b) the self-sustaining nature of the equilibrium points becomes unstable: a small change in x_2 and/or β may trigger a fast development whereby x_1 takes on an entirely different equilibrium value. In such cases, the predictability with regard to x_1 will be low, even if the functional form f_1 were exactly known.

Another example of bifurcation concerns the *nature* of the equilibrium points, which may alter in response to a small parameter change.

Both types of bifurcation (i.e., those related to the number and nature of equilibria) essentially reflect situations of structural instability; they may even occur simultaneously.

A final and different form of bifurcation is related to small changes in the *initial* conditions. A small shift in the initial conditions may force the system to move to a completely different trajectory. This divergence is caused by the shift in the influence of either attracting points (in case of stable equilibria) or repelling points (in case of unstable equilibria). The treatment of unstable system behaviour often requires a probabilistic approach, as we shall demonstrate below.

4. THREE LEVELS OF MODELLING DYNAMICS

Models can be designed and estimated at different levels of aggregation, ranging from macro to micro levels. Each aggregation level imposes certain constraints on model specification and validation, as well as on the conclusions to be inferred from the model's results (see also Blommestein and Nijkamp, 1985). For instance, when a model is specified and estimated as a macro model, one has to be extremely cautious in drawing conclusions about the micro behaviour of economic agents (the so-called problem of 'ecological fallacy').

Each scale of aggregation implies that certain aspects are omitted or are assumed to be constant. For instance, for certain modelling purposes (e.g., short-term forecasts) one may abstract from the explanation of slow moving variables (by using a ceteris paribus clause). In other cases one may be willing to neglect the impact of fast moving processes. For instance, if these processes are assumed to be so fast that the analysis is not distorted, the trajectory of these variables from one equilibrium to another need not be studied precisely (e.g., in a comparative static framework). Clearly, the validity of these simplifying assumptions depends on the differences in time scales and the order of magnitude of the variables involved. However, in a dynamic model fast and slow dynamics may occur simultaneously.

One should note, however, that the rate of change of a variable in a model also usually depends on the variables which are omitted. The behaviour of such variables depends in turn on other variables (whether included in the model or omitted) etc. Consequently, from the viewpoint of specification analysis, it should be observed that a closed set of

equations (i.e., a finite set of equations specified only in terms of agreed-upon variables) only exists in an approximative sense.

There are, however, phenomena for which the distinction according to time scale and size does not provide a useful way of demarcating a closed set of equations. For instance, in the case of structural change processes, we can imagine that a small change at the micro level of a system (e.g., the construction of a road or the introduction of a new production process) may have substantial impacts on the macro level. In these cases "... the difficulty comes from the fact that couplings may exist at all scales from the smallest initial ones to those of the macroscopic level when a system is about to topple over from one stable mode of operating to another" and thus "... a model or theory for describing systems near these critical points must therefore take all these correlations into account in one way or another" (see Courtois, 1985, p. 593).

If the above-mentioned new mode of operation is *qualitatively different* from the original one, one can speak of an evolutionary event (see also Johansson and Nijkamp, 1986): sometimes a small change at the micro level may change the structure of the system. For instance, due to a new road the transportation cost matrix for a whole spatial system may change, or in the case of a new production process new firms, competing with better products, may enter the market. Of course, such new macroscopic structures emerging from microscopic events would in turn have an impact on the structure and the functioning of microscopic mechanisms.

We shall now discuss three levels of change in physical systems as distinguished by Prigogine (1981):

(i) a macro-phenomenological level
(ii) a micro-stochastic level (usually based on Markov processes)
(iii) an approach based on the dynamic laws corresponding to a basic (micro or meso) level.

The fact that we can distinguish a micro level and a macro level (structure) in both physical and social systems may provide a useful analogy. Therefore we will use this distinction to classify types of non-linear dynamic models that have been applied in (regional) economics. Thus, the aim is not to propose unambiguous design and specification principles for such models; such questions are co-determined by the nature of the phenomenon under consideration, the specific research questions posed etc.

(i) Macro-phenomenological level

The variables in the macro-phenomenological approach are (weighted) average values of micro variables whose fluctuations are supposed to have little impact on the first-mentioned variables. Clearly, this assumption is not always warranted; witness the occurrence of bifurcation in real world systems deteriorating the macroscopic description. In case of a bifurcation (which ultimately always stems from the micro level), it is clear that complementary theoretical considerations - not included in the macroscopic viewpoint - are needed in order to adequately analyze the bifurcation process and the way the system reorganizes itself.

(ii) Micro-stochastic level

In the microstochastic approach the micro variables underlying the macro behaviour are explicitly modelled. The factual knowledge regarding their behaviour is limited, so that a probabilistic approach is often followed (e.g., by describing the state transitions of micro variables by Markov processes). By assuming an initial probability distribution for the state of the system, it is then possible to trace the consequences of the micro behaviour and to derive a stationary distribution function. This distribution reflects the fact that the multitude of individual events taking place simultaneously at the micro level of the system

compensate each other statistically, so that they may create a certain macro order which is called a *structure* (cf. als the notion of entropy in a spatial interaction system; see Wilson, 1970).

Under certain conditions the stationary distribution may be multimodal; viz. if - given an initial unimodal distribution and its ensuing trajectory - certain points emerge over time so that the distribution shifts from a unimodal to a multimodal one. Such a transition is accompanied by large fluctuations in the micro variables. It is, in principle, possible to derive approximate mathematical expressions for the growth of such fluctuations proceeding the occurrence of a bifurcation. These fluctuations reflect the existence of a certain ambiguity in the system, as the system may 'choose' between various regimes. Beyond such a bifurcation point the average value of the variables is no longer directly related to the extreme points. Then a multimodal stationary probability distribution results, which indicates that there may be various macro structures that are consistent with the stochastic behaviour of the micro variables. If this micro behaviour is represented by means of a parametrized model, the form of the stationary distribution function may change drastically due to a (small) shift in one of the parameters. As a description of a relatively large system in terms of a probability distribution is often not very meaningful, one usually utilizes mean value equations. It should be noted, however, that in multimodal stationary distributions the relationships between extreme points and mean values become blurred. Fortunately, it can be demonstrated that the stable equilibrium points of the mean value equations correspond to the extreme points of the stationary distribution. When the behaviour of the micro variables depends on the micro state of the system, the mean value equations will be non-linear and there may be multiple stable modes of operation.

(iii) Dynamic laws at a basic level

The third approach provides a description in terms of the dynamic laws operating at a basic level (e.g., the individual trajectories of molecules in a physical system, or the dynamic behaviour of individuals or firms in an economic system). Clearly, the precise demarcation of a basic level implies some arbitrariness. Instead of providing a sharp demarcation criterion for cases (i) and (iii), it is more meaningful to spell out the consequences of *not* using a phenomenological approach; i.e., in what sense does the analysis change if, in one way or another, one takes into account the fluctuation, diversities and feedbacks at the micro level?

Let us assume that the exact dynamic laws governing the behaviour of basic variables are known. Such laws may express stable behaviour, so that neighbouring points are transformed into neighbouring points. Then we may use these laws to analyze the behaviour of the system. However, if these laws reflect unstable behaviour, a problem arises, as then any region in the state space, *whatever its size*, always contains different trajectories that diverge over time. In such cases, even small differences in initial conditions may be amplified. Since a specific limit transition in which processes of a region in the state space are restricted to a point (and hence to a well defined trajectory) is not possible, the description in terms of trajectories breaks down. Then a description in terms of bundles of trajectories becomes relevant. It can be modelled by representing the dynamic equation in stochastic form. This approach, called in the natural sciences the *ensemble* standpoint, is based on a probability aggregate, which is composed of an ensemble of copies of the original system that are consistent with the information assumed about the original system.

The above-mentioned classification of dynamic models is not only relevant for the natural sciences, but also for the social sciences. Various kinds of non-linearities are particularly likely to exist in disaggregate models. Such non-linearities can make the dynamic behaviour of the pertaining model unstable.

In the next section, it shall be demonstrated that various models which in the past years have been developed to describe non-linear evolutions, especially of spatial (urban and regional) socio-economic systems, can be classified by means of these three categories.

5. A CLASSIFICATION OF MAJOR TYPES OF NON-LINEAR DYNAMIC SPATIAL MODELS

Various models have been developed by economists and geographers to describe non-linear dynamic (sometimes irreversible) socio-economic developments in time and space. Instead of providing an exhaustive survey of the literature, we shall present one or two prototype models, which may be regarded as representative for a broad class of models for each class of non-linear dynamic models, discussed in Section 4.

(i) Macro-phenomenological models

Various models in this class are directly or indirectly based on the so-called Volterra-Lotka approach in population dynamics. An example of the macro-phenomenological approach to urban growth and form can be found in Dendrinos (1984) and Dendrinos and Mullally (1985). The central point in this approach is that - despite the complexity of a system at the micro-level - it is possible to gain basic insights into the nature of urban evolution by means of analysing a limited number of strategic macro-level variables. Dendrinos refers to May (1971), who has shown that - in case of random connectance - a system is more likely to be stable when it is small, the elements of the system are weakly connected and the average strength of interaction is low. Although he admits that inter-city linkages are highly non-random, he nevertheless uses this argument (together with the analytical intractability of large non-linear dynamic models) to reject the use of large-scale models for describing urban evolution in the U.S., which has shown remarkable stability over the time-span of a century.

He then introduces the concept of an *effective environment*, which allows him to design a small model for analyzing the income-population dynamics relative to that of the environment of the SMSA's in the U.S. An effective environment is a system's environment, which implies such a normalization of variables that their dynamics can be described by means of the Volterra-Lotka dynamics (or any other non-random dynamic model) which provides theoretical insights and makes empirical verification possible (see Dendrinos, 1985, p. 68).

The standard form of this model is as follows:

$$\begin{aligned} \dot{x}_t &= \alpha(y_t - 1)x_t - \beta x_t^2 \\ \dot{y}_t &= \gamma(\bar{x} - x_t)y_t \end{aligned} \qquad (5.1)$$

where x_t is the relative population size of a metropolitan area (normalized with respect to the total national population size) at time period t, y_t the ratio of urban real per capita income to the prevailing national average during each time period t, and $\bar{x}$ the carrying capacity (in terms of population size) of the metropolitan area concerned.

It has been claimed that this model structure is supported by empirical evidence. The model has to be regarded as a specimen of the macro-phenomenological type, as relative income and population dynamics of an urban area could, in principle, equally well be modelled in terms of disaggregated variables (such as employment in specific industries, the presence of housing, infrastructure etc.).

The insight which the model may provide is rather limited. The behaviour of an urban system can not realistically be explained without consideration of its environment. Despite the normalization used, no functional relationships of the city and its environment are

considered. One of the more interesting issues in regional and urban economics is the question: what are the determinants of the relative carrying capacity of an urban area and what are the driving forces of urban dynamics? The previous analysis, however, provides no answer to these questions. The relative carrying capacity level of population is not an endogenously determined variable in this approach, but is treated as a parameter for which it is claimed that robust estimates are obtainable. The robustness of these estimates is explained by means of the metropolitan areas having sticky ties to their environment. A dynamic theory should be able to explain why this is the case.

One may claim that the above-mentioned model is obtained by implicitly simplifying a more comprehensive model. The validity of explaining a complex phenomenon by means of a few strategically placed macro observations is directly related to the existence of an effective environment, which itself is a somewhat vague concept. It is questionable whether there are many phenomena for which this is possible. In fact, low dimensional models may be useful within a more comprehensive, disaggregated analysis. In the words of Casti (1985, p. 213): " it has been empirically observed in many modelling exercises that the essential behavioral properties of a system which involves interactions of many variables can be captured by centering attention upon a small number of macro-level variables formed, generally, as some (usually non-linear) combination of micro-variables. Usually, the observed macro-variables exhibit the characteristic oscillations, bifurcations, etc., and what is needed is some sort of meso-level theory enabling us to translate back-and-forth between the micro-variables, which we cannot see or know, and the macro-patterns".

(ii) Micro-stochastic models

An example of this type of model can be found in the work of Haag and Weidlich (1984) on interregional migration. They employ synergetic concepts to analyse the dynamics and possible stable modes of operation of a system that describes the distribution of a given population (N) over a number of regions (L). Within their approach the behaviour of the micro variables of the system (i.e., individuals who may or may not migrate to another region) is explicitly considered. It is argued that the heterogeneity of their behaviour impedes a fully deterministic model. Instead, they make a plea for a probabilistic treatment in terms of *transition probabilities* for well defined states. These transition probabilities provide a Markov chain. They may be modelled by means of a parametrized model. The parameters in such a model are called trend parameters. The resulting stochastic theory describes the system in terms of a probability distribution defined over the possible states of the system.

By means of the transition probabilities it is possible to link (static) theoretical considerations to dynamics, and by means of the so-called *master equation* - which describes the evolution of the probability distribution over time - a link is provided between the micro- and macro-level. Thus an elegant framework for analysing the dynamics of complex systems results. Once the initial conditions and the specification of the transition probabilities are given, the behaviour of the system is completely determined by the numerical values of the trend parameters. These trend parameters are then estimated on the basis of empirical data. In a more extensive analysis they may be explained by socio-economic factors and in this way the model becomes more appropriate for prediction purposes (see Annex 2 for more details).

In the case that the transition probabilities are functions of the macro-state of the system, there exists a feedback from the macro- to the micro-level. This feedback may lead to a multimodal stationary distribution function, the shape of which may change drastically under certain critical changes of the trend parameters (bifurcation). The states of the system corresponding to the maxima of the stationary distribution are to be interpreted as stationary end-states in which the spatial interaction system has attained a stable mode of operation. In the case of a multimodal distribution there are several

stationary end-states depending on the initial conditions under which it was attained. Under the condition of 'detailed balance' - that is, local balance of all probability fluxes - it is possible to derive explicitly the stationary distribution function. An often analytically more tractable but less informative representation is obtained by means of value equations. These deterministic equations describe how the mean values of the number of people living in the different regions change as the probability distribution evolves over time. In formula:

$$\frac{d\bar{n}_i}{dt} = \sum_{\bar{n}_i} n_i \frac{dp(n;t)}{dt} \tag{5.2}$$

where :

$n' \equiv (n_1,..n_i,..n_j,...n_L)$, a vector consisting of the number of people living in the various regions.

$\bar{n}_i =$ the mean value of the number of people living in region i.

$\frac{dp(n;t)}{dt}$ = the time derivation of the distribution function concerned.

By an approximation which is valid as long as the distribution remains narrow and unimodal, a closed (self contained) set of L differential equations is obtained.[1] These equations are non-linear if the individual transition probabilities are functions of n. The stationary points of the mean value equations correspond to the states at which the stationary distribution attains its maxima. Possible bifurcation phenomena are also reflected in these mean value equations. The stationary points which are finally attained by the mean value equation are dependent on the initial conditions.

(iii) Models based on dynamic laws at a basic level

An example of the disaggregated approach belonging to the third class is the work of Allen and Sanglier (1981). Their work is part of the investigation of 'self-organizing' phenomena in natural and social systems. In this approach the interactions or geographically distributed sites can be analysed. This leads to dynamic equations in which various types of non-linearities are present, and consequently the dynamics expressed in these models may reflect an unstable behaviour.

A typical example of these equations reads:

$$\frac{dx_i}{dt} = bx_i\left(J_i^o + \sum_k J_i^k - x_i\right) - mx_i + \tau\left[\sum_{j\neq i}\left\{x_j^2 \exp(-\beta d_{ij}) - x_i^2 \exp(-\beta d_{ij})\right\}\right] \tag{5.3}$$

where

x_i - population of site i

J_i^o - basic 'carrying capacity' of site i

J_i^k - number of jobs in activity k at site i

d_{ij} - distance between site i and site j

The parameters b and m reflect the demographic change (birth and death rates) as well as the immobility of the population in residential relocation under pressure from the distribution of available employment. The last term, consisting of a weighted sum of the

[1] The possible development in case of non-linearity into a multimodal distribution is an *extremely slow process* which takes place after the distribution has been centred around one of the (ultimate) maxima.

squared number of people living in the different site, expresses the influence of congestion effects.

The specification of (5.3), which consists of an equilibrium condition and a dynamic adjustment process, is hardly motivated. The dynamics employed is inter alia used in biology (Volterra-Lotka). There it is reasonable to assume, under certain circumstances, that $x_1.x_2$ is a suitable proxy (model) for the number of prey-predator interactions between species 1 and 2. It is questionable whether these types of dynamics are appropriate for modelling socio-economic processes. In any case, they imply a type of behaviour that is not motivated.

Allen et al. consider spatial structures as being far from equilibrium in a thermodynamic sense and consequently they presuppose that flows of matter, persons and energy lead to a maintenance of this disequilibrium situation. The fact that a social system is open is incorporated in the model by two types of stochastic processes which may influence the simulation results in a non-trivial way. One is the random introduction of economic activity at various time intervals at all locations; if certain exogenous boundary conditions are met, the activity will develop, otherwise it will not survive. The second type concerns the perturbation of exogenous parameters and the deviations of behaviour of an individual agent from an average aggregate performance level. The authors are not so interested in making exact predictions, but rather in illustrating the consequences of fluctuations and non-linearities. Their explanation is not as 'strong' as the usual 'causal' explanations of classical physics. "It is through the action of elements not explicitly contained in the equations (fluctuations or historical 'accidents') that the choices are in fact made at the various bifurcation points that occur during the evolution of any particular system" and "the spatial organization of a system does not result uniquely and necessarily from the 'economic and social laws' enshrined in the equations, but also represents a 'memory' of particular specific deviations from these average behaviours" (Allen and Sanglier, p. 168).

Another example of the approach based on 'basic' dynamic laws is the type of models employed by Wilson et al. (see for example Wilson, 1981). They have used their models to analyse among other things the urban retail structure and the residential structure. Here we shall take the urban retail structure model as an example. In this model two types of agents are distinguished: there are consumers who respond via their buying behaviour to a given spatial distribution of shopping centres and there are entrepreneurs who determine their investments and consequently the urban retail structure in response to revenue generated in the different shopping centres. The buying behaviour of the consumers is modelled by means of a measure of the attractiveness of the different centres and the costs of travel. The attractiveness of the centre is assumed to be a function of the size of the centre. This size itself is determined by the investment behaviour of the entrepreneurs.

The total revenue (expenditure on consumption goods) attracted to centre j is:

$$D_j = \sum_i \frac{e_i p_i W_j^{\alpha} e^{-\beta c_{ij}}}{\sum_j W_j^{\alpha} e^{-\beta c_{ij}}} \qquad (5.4)$$

where:

e_i = the average per capita expanditure on shopping goods by the residents of zone i

P_i = the population of zone i

W_j = the size of centre j

c_{ij} = the cost of travel from i to j

α,β = parameters

Further, it is postulated that the cost of employing a centre is proportional to its size:

$$c_j = kW_j \tag{5.5}$$

where k is a suitable constant. To analyse the development of the retail structure it is assumed that the entrepreneurs will expand their facility if there are positive profits and that the facility will be reduced if they make a loss.

So the following equilibrium condition is postulated:

$$\sum_i \frac{e_i p_i W_j^{\alpha} e^{-\beta c_{ij}}}{\sum_j W_j^{\alpha} e^{-\beta c_{ij}}} = kW_j \tag{5.6}$$

$$\dot{W}_j = F\left[\sum_i \frac{e_i p_i W_j^{\alpha} e^{-\beta c_{ij}}}{\sum_j W_j^{\alpha} e^{-\beta c_{ij}}} - kW_j\right] \tag{5.7}$$

where F is a function such that $F[0] = 0$ and $F'[x] > 0$, $\forall$ x. It is worth noting that the specification of the F-function - in principle to be based on theoretical considerations - influences the nature of the equilibrium points and hence the dynamic trajectory. The model is highly non-linear in the W_j variables. (5.6) models the distribution of the expenditure over the different centres. The denominator in (5.6) serves to make this distribution consistent. It leads to non-linearities and the possibility of multiple equilibria. The dynamic behaviour of the model may exhibit bifurcation properties that are not easily investigated analytically. This will be even more likely when the model is disaggregated and the attractiveness factors themselves become non-linear functions.

After having briefly considered the various prototypes of models belonging to the three above-mentioned classes, we shall draw some more specific conclusions regarding the modelling of non-linear dynamic (spatial) systems in the final section.

6. CONCLUSION

Non-linear dynamic systems are appropriate tools for modelling discontinuous and structural changes in real-world systems.

A non-linear dynamic system may, for ranges of numerical values of its parameters, have multiple equilibrium points. So the equilibrium which is finally attained may be path-dependent. Non-linearity therefore requires a dynamic analysis and often also a disequilibrium approach. This poses problems. As Koopmans (1957) observed 'until we succeed in specifying fruitful assumptions for the behaviour in an uncertain and changing economic environment, we shall continue to be groping for the proper tools of reasoning'. We still lack such a fruitful theory of behaviour in disequilibrium.

The non-linearity may have its origin in non-linear individual behavioural relationships or it may be caused by aggregation of linear ones. However, theoretical knowledge and empirical information concerning functional forms in economics is very limited and the class of non-linear functions is very wide. Conscientious econometric research may be helpful, but this still requires basic assumptions which may be hard to motivate and will require an enormous effort to test empirically. Comparing different empirical models of a certain phenomenon that are based on different functional forms is far from easy and is not guaranteed to lead to conclusive results. It therefore seems an exacting endeavour to describe specific non-linear phenomena by means of a complete set of well-motivated non-linear dynamic laws.

Thereby arises a dilemma. On the one hand there are empirical phenomena that cannot be explained by linear models, while on the other hand the knowledge related to behaviour in disequilibrium and functional forms is often too limited to warrant the specification of a set of non-linear dynamic laws.

A completely different and in our view promising way of arriving at a non-linear system is the synergetic approach. Synergetics is defined as the science of collective static or dynamic phenomena in closed or open multi-component systems with "cooperative" interactions occurring between the units of the system (Weidlich and Haag, 1983, p. 1). The non-linearity follows from the assumption that the behaviour of the micro-units (the components) of the system is dependent on its macro-state. The macro-state of the system is at the same time influenced by the individual behaviour. There exists a cyclic coupling between causes and effects which may lead to multiple self-sustaining states (structures). The non-linearity enters in a 'natural' way and not by means of an often arbitrary functional form. Since in the synergetic approach the micro- and macro-level are modelled simultaneously it is suited for the analysis of structural change.

The behaviour of the micro-units is modelled by means of transition probabilities which define a Markov process. Therefore a specific functional form has to be chosen. Modelling the transition probabilities, however, poses a rather *well ordered problem*, relatively easily accessible to empirical investigation, as opposed to *the all embracing nature of the specification of dynamic laws*. Within the synergetic approach it is not necessary to postulate equilibrium conditions and disequilibrium behaviour. Once the transition probabilities are modelled, the dynamics and equilibria follow in a logical way.

The synergetic approach leads to a description in terms of a distribution function over the state space. This description can be shown to be consistent with a description in terms of stochastic dynamic equations for the relevant variables (see Weidlich and Haag, 1983). If the deterministic part of these equations exhibits unstable behaviour the latter description is rather limited. Simulations of the dynamic trajectories will lead to a wide spectrum of results. These results reflect the above-mentioned distribution function. The description in terms of the distribution function is to be preferred since it is more complete. The intractability of the distribution function of somewhat larger systems is, however, a serious handicap. One should then resort to the mean value equations as these provide (often adequate) information about the distribution function in an elegant and concise way.

In retrospect: the foregoing observations lead us to the conclusion that the strength of the master equation is its ability to account for synergetic effects in social systems in an appropriate manner. Further empirical research, including an examination of the actual properties of the mean value equations for real-world systems, is no doubt warranted.

Annex 1. Stability Analysis by Means of Eigenvalues

In this Annex, the way in which equilibrium points x^* can be dealt with on the basis of an eigenvalue approach of the matrix $DF(x^*)$ will be outlined.

The linearized vector field that is close enough to the equilibrium point x^* describes its stability properties. If at least one of the eigenvalues has a non-zero imaginary part, the convergence or divergence will exhibit an oscillatory pattern. Clearly, when $F(\cdot)$ is a *linear* function, there is at most one equilibrium point. In the latter case, $DF(\cdot)$ is constant, so that the local behaviour and the global behaviour of the system coincide; besides, the qualitative characteristics of a solution will be independent of the initial conditions. As the linear case implies that $DF(x^*)$ will generally be a function of the parameters, the equilibrium will change due to a shift in parameters. However, the location of the equilibrium point will not change dramatically in response to a small parameter shift.

On the other hand, if $F(\cdot)$ is a *non-linear* function, several equilibrium points may occur; their number depends inter alia on the numerical values of the parameters. $DF(\cdot)$ is

then in general a function of the parameters and of x. Such a case of multiple equilibria of non-linear models implies that sometimes the *location*, *existence* and *character* of equilibrium points may drastically change in response to a small change in one of the parameters.

Various examples of non-linear dynamic models can be found in regional economics (see also Section 5). Several of these models incorporate a static equilibrium condition in a dynamic framework. This equilibrium condition determines the *location* of the equilibrium points, while their *character* is also dependent on the dynamic framework. This approach bears a similarity to Samuelson's correspondence principle stating that a fruitful application of comparative static methods often presupposes a theory of dynamics (see Samuelson, 1948).

In conclusion, instead of specifying a priori and in an uncritical way a specific type of dynamics (e.g., Volterra-Lotka dynamics, May dynamics) for economic systems, it is preferable to give sufficient attention to the design of models for dynamic adjustment processes of the dynamic system at hand.

Annex 2. The Master Equation Approach

In this Annex synergetic concepts employed by Haag and Weidlich (1984) will be described. The behaviour of the (migrating) individuals is modelled by means of individual transition probabilities P_{ji} which describe the probability that an individual will migrate from region i to region j ($i \neq j$). With respect to these probabilities the Markov assumption is made. Thus the probability than an individual migrates from i to j in a given time interval is independent of its behaviour preceeding that time interval. The probabilities, which link static concepts to dynamics, are modelled as follows:

$$P_{ji}(n_j,n_i) = v \exp[f_j(n_j+1) - f_i(n_i)] \quad i \neq j \tag{I}$$

where:

P_{ji} = individual transition probability for a transition from region i to region j
v = global mobility parameter determining the time scale of the migration process
n_k = number of people living in region k
$f_k(n_k)$ = utility of an individual living in region k which has a population n_k; the trend parameter which enters this function may be explained by socio-economic factors within the context of a regression model.

These individual transition probabilities enter into the expression for transition 'probabilities' between so-called socio-configurations. These socio-configurations are the possible distributions of the given N individuals over the given L regions. A typical socio-configuration can be described by the vector n containing a given number of individuals living in each region. The following model for the transition $n = (n_1,...,n_j,...n_i,...n_L) \rightarrow n^{ji} = (n_1,...,n_j+1,...,(n_i-1),...,n_L)$ results:

$$W_{ji}[n] = n_i P_{ji}(n_j,n_i). \tag{II}$$

All the n_i members contribute to the probability P_{ji}. (II) enters into the master equation, which describes the evolution of the probability that the system is in one of the $\binom{N}{L}$ socio-configurations.

$$\frac{dP(n;t)}{dt} = \sum_{i,j=1}^{L} \{W_{ji}[n^{ij}]P([n^{ij};t) - W_{ji}[n]P(n,t)\}. \qquad \text{(III)}$$

The first and second term within the summation express a probability flux that goes into, respectively out of, the socio-configuration n. Summation gives the net result. Since there are $\binom{N}{L}$ socio-configurations the master equation is a system of $\binom{N}{L}$ coupled linear differential equations. It can be proved that the distribution function P(n;t) finally becomes time independent. In the case of 'detailed balance' (i.e. $W_{ji}[n^{ij}]P(n^{ij};t) = W_{ji}[n]P(n;t) \forall i,j,i \neq j$) the stationary distribution can be obtained. A less informative, but often more tractible description is obtained in terms of mean value equations (see page 181). The exact equations of motion for the mean values, which may be derived using (III), read:

$$\frac{d\bar{n}_j}{dt} = \sum_{i=1}^{L} \overline{W_{ji}[n;t]} - \sum_{i=1}^{L} \overline{W_{ij}[n;t]} \qquad j=1,....,L \qquad \text{(IV)}$$

where $\overline{W_{ji}[n;t]} = \sum_{n} W_{ji}[n;t]P(n;t)$.

Employing the approximation $\overline{W_{ji}[n,t]} = W_{ji}[\bar{n};t]$, which is valid as long as the probability distribution remains narrow and unimodal (see footnote 1, page 181), we arrive at the self-contained set of coupled-differential equations:

$$\frac{d\bar{n}_j}{dt} = \sum_{i=1}^{L} W_{ji}[\bar{n};t] - \sum_{i=1}^{L} W_{ij}[\bar{n};t] \qquad j=1,...,L. \qquad \text{(V)}$$

Since P_{ji} is a function of n, (V) is non-linear and may have distinct stationary points. For the relation between these mean value equations and the distribution function see page 181.

REFERENCES

Allen, P.M. and M.Sanglier, 1981, "Urban Evolution, Self-organization, and Decision-making", *Environment and Planning A*, 13:167-183.

Arrow, K. and F. Hahn, 1971, *General Competitive Analysis*, North-Holland, Amsterdam.

Beckmann, M.J. and T. Puu, 1985, *Spatial Economics: Density, Potential and Flow*, North-Holland, Amsterdam, New York, Oxford.

Blommestein, H.J. and P. Nijkamp, 1985, "Testing the Spatial Scale and the Dynamic Structure in Regional Models", *Journal of Regional Science*, vol 29.

Brouwer, R. and P. Nijkamp, 1985, "Qualitative Structure Analysis of Complex Systems", in P. Nijkamp, H. Leitner and N. Wrigley, (eds.), *Measuring the Unmeasurable*, Martinus Nijhoff Publishers, Dordrecht, Boston, Lancaster.

Casti, J.L., 1985, "Simple Models, Catastrophes and Cycles", RR-85-2, International Institute for Applied Systems Analysis.

Courtois, P.J., 1985, "On Time and Space Decomposition of Complex Structures", *Communications of the ACM*, 28:590-603.

Dendrinos, D.S., 1984, "Madrid's Aggregate Growth Pattern: A Note on the Evidence Regarding the Urban Volterra-Lotka Model", *Sistemi Urbani*, vol. VI, no. 3.

Dendrinos, D. and H. Mullally, 1985, "Urban Evolution", *Studies in the Mathematical Ecology of Cities*, Oxford University Press, New York.

Frisch, R., 1935-1936, "On the Notion of Equilibrium and Disequilibrium", *Review of Economic Studies, III*, pp. 100-106.

Griffith, D.A. and A.C. Lea, (eds.), 1983, *Evolving Geographical Structures*, Martinus Nijhoff Publishers, The Hague, Boston, Lancaster.

Haag, G. and W. Weidlich, 1984, "A Stochastic Theory of Interregional Migration", *Geographical Analysis*, vol 16, 4:331-357.

Hannan, M.T. and N.B. Tuma, 1985, "Dynamic Analysis of Qualitative Variables: Applications to Organizational Demography", in Nijkamp et al., (eds.), *Measuring the Unmeasurable*, Martinus Nijhoff Publishers, Dordrecht, Boston, Lancaster.

Johansson, B. and P. Nijkamp, 1986, "Analysis of Episodes in Urban Event Histories" in L.Van de Berg, L. Burns and L.H. Klaassen, (eds.), *Spatial Cycles*; Gower, Aldershof, U.K.

Koopmans, T.C., 1957, *Three Essays on the State of Economic Science*, McGraw-Hill Book Company, New York, Toronto, London.

Li, T.Y. and J.A. Yorke, 1982, "Period Three Implies Chaos", *American Mathematical Monthly*, 82, 10:985-992.

May, R.M., 1971, "Stability in Multi-Species Community Models", *Mathematical Biosciences*, 12:59-79.

May, R.M., 1974, "Biological Populations with Nonoverlapping Generations: Stable Points, Stable Cycles and Chaos", *Science*, 186:645-647.

Nicolis, G. and I. Prigogine, 1977, *Self-Organization in Nonequilibrium Systems, from Dissipative Structures to Order through Fluctuations*, John Wiley and Sons, New York.

Pimm, S.L., 1982, *Food Webs*, Chapman and Hall, London.

Prigogine, I., 1981, "Time, Irreversibility, and Randomness", in E. Jantsch, (ed.), *The Evolutionary Vision, Towards a Unifying Paradigm of Physical, Biological and Sociocultural Evolution*, Westview Press., Inc., Boulder, Colorado.

Rijk, F.J.A. and A.C.F. Vorst, 1982, "Equilibrium Points in an Urban Retail Model and their Connection with Dynamical Systems", report 8214/M, Erasmus University, Rotterdam.

Samuelson, P.A., 1948, *Foundation of Economic Analysis*, Harvard University Press, Cambridge.

Thom, R., 1975, *Structural Stability and Morphogenesis*, W.A. Benjamin, Luc, Mass.

Varian, H.R., 1981, "Dynamical Systems with Applications to Economics", in K.J. Arrow, and M. Tutriligatu, (eds.), *Handbook of Mathematical Economics*, volume 1, North-Holland Publishing Company, Amsterdam, New York, Oxford.

Weidlich, W. and G. Haag, 1983, *Concepts and Models of a Quantitative Sociology: the Dynamics of Interacting Populations*, Springer Verlag, Berlin, Heidelberg, New York.

Wilson, A.G., 1970, *Entropy in Urban and Regional Modeling*, Pion, London.

Wilson, A.G., 1981, *Catastrophe Theory and Bifurcation*, Croom Helm, London.

PART C

URBAN AND REGIONAL INFRASTRUCTURE

Advances in Spatial Theory and Dynamics
A.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 13

Dynamics of County Growth

E.S. Mills and G. Carlino

1. INTRODUCTION

The 1970's witnessed continuing and, in some respects, accelerating movements of people and jobs within the U.S. The movement from metropolitan central cities to suburbs continued unabated while there was also evidence of movement of both people and jobs from metropolitan to non-metropolitan areas, reversing a long standing trend. Regionally, people and jobs moved from the eastern, northeastern and north central to southern and western parts of the country. (For documentation of these trends see Carlino (1985) and Mills (1986)). This regional movement is known popularly and inadequately as movement from the frostbelt to the sunbelt.

Why have these movements occurred? Have metropolitan areas become inherently unattractive or has migration from them resulted from controllable phenomena? Are climate and other amenities mainly responsible for the move to the sunbelt? Do people follow jobs or vice versa? How have federal, state and local government actions affected the movements? Carlino (1985) analyzes the demetropolitanization of manufacturing employment in terms of the decreasing importance of agglomeration economies. Mills (1986) analyzes the determinants of relative growth of metropolitan central cities and suburbs.

This paper presents an analysis of growth in a small area during the 1970's, employing an explicit model and a large comprehensive data set. We believe the approach adopted here has the potential to shed light on all the questions posed in the previous paragraph.

2. THEORETICAL MODEL

We envisage a conventional general equilibrium model in which both producers and consumers are geographically mobile. Consumers maximize utility which depends on purchased goods and services, on location relative to work places and on spatially varying non-market amenities. A conventional budget constraint equates income to the sum of spending on goods and services. Local taxes reduce consumption expenditures and government services and other amenities appear in utility functions.

Firms produce goods and services, buying inputs and selling outputs in competitive markets. Production costs vary by location because of regional comparative advantage, regional variation in labor supply (amenities affect labor supplied at given wage rates),

and perhaps because of spatial variation in other government actions - land use controls, state and local taxes, right-to-work laws, etc.

Among the endogenous variables in such a general equilibrium model are prices and quantities of inputs and outputs by location. No data exist for most such variables for small geographical areas. Data exist by county for population and most employment, and these related variables are of great government and private concern. We assume that other endogenous variables have been solved out of the general equilibrium model and analyze joint determination of population and employment in small areas.

Specifically, we assume that the model just outlined can be solved to give

$$E^* = f(P,S) \tag{1}$$

and

$$P^* = g(E,S) \tag{2}$$

where E and P are employment and population, S is a vector of exogenous variables that affect E and P, and the asterisks indicate equilibrium values. All variables refer to counties, but subscripts are suppressed. (1) and (2) assume that equilibrium employment and population depend on actual population and employment and on the variables included in S.

County population and employment are assumed to adjust to equilibrium with distributed lags:

$$E = E_{-1} + \lambda_E(E^* - E_{-1}) \tag{3}$$

and

$$P = P_{-1} + \lambda_P(P^* - P_{-1}) \tag{4}$$

where subscript -1 refers to the value of the indicated variable lagged one period - a decade in our data - and λ_E and λ_P are speed of adjustment coefficients.

Substituting (1) and (2) for E^* and P^* in (3) and (4), and rearranging terms, gives

$$\Delta E = \lambda_E f(P,S) - \lambda_E E_{-1} \tag{5}$$

and

$$\Delta P = \lambda_P g(E,S) - \lambda_P P_{-1}. \tag{6}$$

(5) and (6) contain only observable variables. Those forms were chosen in preference to alternatives with $\Delta E/E_{-1}$ and $\Delta P/P_{-1}$ as dependent variables on statistical grounds and after some experimentation.

Having estimated (5) and (6), long run or steady state effects of changes in exogenous variables can be calculated by setting $E = E_{-1}$ and $P = P_{-1}$ in the estimated equations and solving the resulting simultaneous equations to obtain long run equilibrium values $\bar{E}$ and $\bar{P}$ of E and P, where

$$\bar{E} = F(S) \tag{7}$$

and

$$\bar{P} = G(S). \tag{8}$$

Derivatives of (7) and (8) show estimated long run effects of components of S on employment and population.

3. DATA

The basic data set employed in this study is the tape for County Business Patterns. The tape contains the 1970 and 1980 census of population counts for each county. It also contains counts of 1969 and 1979 employment by place of employment.[1] The tape also contains a variety of other census variables that are candidates for inclusion in the S vector. We added several variables from other sources.

Inevitably, judgment and experimentation are entailed in specifying the S vector. Previous studies have analyzed effects of such variables as taxes, crime, racial composition, and the adequacy of the transportation system on the location of either population or employment, but rarely both. We used earlier studies to attempt to distinguish the variables which are likely to be important components of S. All the experiments we tried are reported in the following section. We were inevitably limited by the set of variables available by county. To make allowance for the effects of unavailable variables, we included dummy variables for the 9 census regions. We also included dummy variables for counties which include metropolitan central cities, for counties that are adjacent to metropolitan areas and for counties that are neither metropolitan nor adjacent to metropolitan areas.

In the U.S. context, racial mix of county population is a natural candidate for inclusion in the S vectors. We have included the percentage of the population that is black. We experimented with several measures of transportation adequacy within counties. By far the most important variable turned out to be miles of interstate highway. Taxes vary by state and county in the United States, because of the autonomy of state and local governments. Total local taxes per capita is a natural variable to include. The difficulty is that we lack adequate measures of the quality and quantity of government services produced. Total income is an important variable. High incomes attract both people and businesses, so we included income per family within each county. A high crime rate presumably deters people and businesses from locating in an area, so we included the value of the index of major crimes for the county. Unionization may deter businesses from locating in a county, but should not affect population. The percentage of the labor force that belongs to unions is available only by State, and we included it at that level. State and local governments vary as to their willingness to issue tax free development bonds. In some states, it is a major development policy, so we included the total of such bonds outstanding in the state in which each county is located.

Counties vary greatly in size. Most of the variables just discussed are normalized by county population, but E and P should be normalized by the county's land area. Therefore, E and P are expressed as employment and population per square mile of land area in the county in all the estimates reported in the next section.

Almost any candidate variable for inclusion in S is likely not only to affect E and P but also to be affected by E and P to some extent. It would require a large simultaneous equation model to take account of all likely interactions. No such model is available. To minimize such interactions, we have chosen S variables for 1970, where ΔE and ΔP

[1]County Business Patterns employment data consistently covered nearly 90 percent of total employment during the years included in this study. Population data reflect residents in a county and employment data reflects jobs in a county, but it is not known how many employees work in their county of residence, except for those in large metropolitan areas.

reflect the changes in 1969-79 and 1970-1980 respectively. The only exception is that the development bond variable is available only for 1981. However, the included variable is the stock of outstanding bonds in 1981, and they were mostly issued during the 1970's, so the stock should show effects on the dependent variable, if effects exists.

Finally, we separated out manufacturing employment for separate estimation. Thus, (5) and (6) were estimated once for total employment included in the County Business Patterns tape and a second time for manufacturing employment. Thus, two sets of two equations each are reported in the next section.

4. STATISTICAL ESTIMATES

Our preferred estimates of (5) and (6) for total employment and population are provided in Table 1. Not surprisingly, population and employment are highly interactive, a large population stimulating employment growth and numerous employment opportunities stimulating population growth. Also, not surprisingly, the effect of initial employment on employment growth and the effect of initial population on population growth is large and significant. Our estimates indicate that employment adjusts to disequilibrium somewhat more slowly than population. Estimated adjustment coefficients of .14 and .16 per decade indicate the importance of disequilibrium adjustments in the data and the danger of assuming that population and employment are in equilibrium in any given year.

The estimates indicate a mild deterrent effect of a high percentage of black residents on population growth. The coefficient is not significant at usual significance levels. The indication may be that racial tensions have lessened during the 1970's. Alternatively, it may indicate that blacks represented a large propotion of the migrants during the 1970's and they are not deterred from migrating to counties with large racial minorities. The coefficient representing the percentage of blacks in the employment growth equation is equally small and insignificant, but positive. A large percentage of black residents probably correlates with low wages, which may attract employment.

A high density of interstate highways attracts both population and employment growth. The indication here is that location of the interstate highway system has had a large effect on population and employment movements. Other highway variables, such as the density of limited access highways, showed much less effect than the interstate highway variable. High local taxes per capita deter population growth suggesting that, at the margin, people attach little value to any additional local government services that higher taxes may finance. Again, the significance level is low. The coefficient of taxes in the employment growth equation was even smaller and less significant, and the variable has been omitted from the employment growth regression reported in Table 1. The insignificance of local taxes in affecting employment location confirms findings in large numbers of previous studies by public finance specialists.

Family income has a large and highly significant positive coefficient in both equations. In the population growth equation, high family income presumably stands for prestigious residential neighborhoods. In the employment growth equation, high family income presumably indicates high demand for locally produced commodities and services. There is no indication in our analysis that stringent zoning in high income counties deters employment growth. The reason probably is that stringent zoning mostly pertains to local government jurisdictions, which cover smaller areas than counties.

Unionization deters employment growth, but the coefficient is not significant at usual significance levels. As should be expected, our analysis showed that unionization has no effect on population growth, for given employment levels.

A high crime rate has a small and insignificant deterrent effect on population growth. The analysis suggested no measurable direct effect of crime on employment, so the variable is excluded from the regression reported in Table 1. These results are surprising and we have no explanation for them. In our sample, the crime rate has a small negative correlation with the percentage of black residents ($r = -.05$).

Our estimates indicate a strong preference by the population for central city over suburban and for suburban over nonmetropolitan counties of residence. But there is little to distinguish preferences between adjacent and non-adjacent counties. The strong suggestion in these results is that population movements from central cities and, to a lesser extent, from metropolitan areas result from values of variables captured in our analysis, not from disamenities such as high population density, pollution or congestion that characterize central cities and metropolitan areas. Perhaps even more surprising, employment shows a preference for central cities over other locations. The dummy coefficients do not suggest significant differences among suburbs, adjacent counties and non-adjacent counties in attracting employment.

Table 1 Structural Equations for Total Employment and Population Growth

	Employment Growth	Population Growth
Intercept	$-0.541(10)^{-1}$	$-0.521(10)^{-1}$
	(-4.334)	(-4.339)
Population	$0.111(10)^{-1}$	-
	(6.424)	-
Employment	-	$0.693(10)^{-1}$
	-	(25.160)
Lagged Population	-	-0.164
	-	(-104.807)
Lagged Employment	-0.135	-
	(60.224)	-
Percent Black	$0.414(10)^{-4}$	$-0.449(10)^{-4}$
	(0.342)	(-0.362)
Interstate Highway Density	0.246	0.111
	(6.568)	(2.752)
Local Taxes per Capita	-	$-0.266(10)^{-4}$
	-	(-1.393)
Family Income	$0.834(10)^{-5}$	$0.148(10)^{-4}$
	(7.165)	(11.087)
Crime Rate	-	$-0.383(10)^{-6}$
	-	(-0.402)
Percent Union	$-0.165(10)^{-3}$	-
	(-0.716)	-

Development Bonds	$0.850(10)^{-8}$ (0.230)	- -
Central City Dummy	$0.450(10)^{-1}$ (7.556)	$.325(10)^{-1}$ (4.839)
Ajacent Dummy	$0.143(10)^{-2}$ (0.284)	$-0.223(10)^{-1}$ (-4.170)
Non-adjacent Dummy	$0.586(10)^{-2}$ (1.120)	$-0.164(10)^{-1}$ (-2.938)
New England	$-0.583(10)^{-2}$ (0.537)	$-0.194(10)^{-1}$ (-1.661)
Mid-Atlantic	$-0.225(10)^{-1}$ (-2.457)	$-0.720(10)^{-3}$ (-0.087)
Eastern Plains	$-0.135(10)^{-1}$ (1.962)	$-0.366(10)^{-1}$ (-6.033)
Western Plains	$-0.887(10)^{-2}$ (-1.684)	$-0.305(10)^{-1}$ (-5.157)
Eastern South Central	$0.423(10)^{-2}$ (0.791)	$0.883(10)^{-4}$ (0.016)
Western South Central	$-0.704(10)^{-3}$ (-0.139)	$-0.823(10)^{-2}$ (-1.541)
Rocky Mountain	$-0.119(10)^{-1}$ (-1.837)	$-0.352(10)^{-1}$ (-4.991)
Pacific	$+0.119(10)^{-1}$ (1.367)	$-0.192(10)^{-1}$ (-2.152)
R^2	0.887	0.858

The regional dummies in the population growth equation do not support the notion that climate and other amenities are the causes of migration to sunbelt counties. The eastern and western plains regional dummies have large negative coefficients and these are frostbelt regions. But New England and the mid-Atlantic dummies have insignificantly negative coefficients. Surprisingly, the Rocky Mountain and Pacific regions, both destinations of net migration, have significant negative coefficients. The employment growth regression shows somewhat stronger frostbelt-sunbelt effects. The mid-Atlantic and two plains regions all have negative coefficients that are significant at the 95 percent level. If the reason were fuel costs, New England should also have a significantly negative coefficient. Our guess is that it represents stronger anti-business attitudes of governments in these regions than elsewhere. The significantly negative coefficient of the Rocky Mountain dummy could represent the same phenomenon or, more likely, adverse topography.

Finally, it should be pointed out that the R^2's are extremely large for such cross-sectional regressions.

Table 2 presents estimates of (5) and (6), employing the same S vector as in Table 1, but interpreting E to be manufacturing employment. Because of disclosure rules, manufacturing employment is available for fewer counties than is total employment, so the sample size is about 2600 in Table 2, whereas it is about 3000 in Table 1.

In Table 2, the population growth regression coefficients are remarkably similar to corresponding coefficients in Table 1. The greatest difference is that the small deterrent effect of high local taxes found in Table 1 disappears in Table 2.

Table 2 Structural Equations for Manufacturing Employment and Population Growth

	Manufacturing Employment Growth	Population Growth
Intercept	-0.021 (-7.260)	$-0.570(10)^{-1}$ (-4.040)
Population	$-0.121(10)^{-2}$ (-3.143)	- -
Manufacturing Employment	- -	-0.355 (19.916)
Lagged Population	- -	-0.165 (-89.577)
Lagged Manufacturing Employment	-0.323 (-153.683)	- -
Percent Black	$0.634(10)^{-4}$ (2.354)	$-0.109(10)^{-3}$ (-0.760)
Interstate Highway Density	$0.728(10)^{-1}$ (9.474)	0.201 (4.526)
Local Taxes Per Capita	- -	$0.395(10)^{-6}$ (-0.146)
Family Income	$0.293(10)^{-5}$ (11.085)	$0.149(10)^{-4}$ (9.478)
Crime Rate	- -	$-0.221(10)^{-6}$ (-0.200)
Percent Union	$-0.211(10)^{-4}$ (-0.407)	- -
Development Bonds	$-0.711(10)^{-6}$ (-0.891)	- -
Central City Dummy	$0.122(10)^{-1}$ (9.913)	$0.221(10)^{-1}$ (2.988)
Ajacent Dummy	$0.130(10)^{-2}$ (1.195)	$-0.205(10)^{-1}$ (-3.343)
Non-adjacent Dummy	$0.258(10)^{-2}$ (2.253)	$-0.148(10)^{-1}$ (2.309)
New England	$-0.669(10)^{-3}$ (-0.298)	$-0.274(10)^{-1}$ (-2.082)
Mid-Atlantic	$-0.192(10)^{-2}$ (-0.988)	$-14.075(10)^{-1}$ (-1.515)
Eastern Plains	$-0.140(10)^{-2}$ (-0.935)	$-0.453(10)^{-1}$ (-6.534)

Western Plains	$-0.203(10)^{-2}$ (-1.710)	$-0.381(10)^{-1}$ (-5.472)
East South Central	$0.272(10)^{-2}$ (2.365)	$-0.111(10)^{-2}$ (-0.176)
West South Central	$-0.149(10)^{-3}$ (-0.134)	$-0.767(10)^{-2}$ (-1.229)
Rocky Mountain	$-0.506(10)^{-2}$ (-3.410)	$-0.399(10)^{-1}$ (-4.735)
Pacific	$-0.225(10)^{-2}$ (-1.208)	$-0.195(10)^{-1}$ (-1.902)
R^2	0.989	0.853

In the manufacturing employment regression in Table 2, the most surprising finding is that population has a significantly negative sign. Presumably high population density implies land use controls or high land values, both of which deter manufacturing firms. Probably more important, manufacturing production often has only small and specialized labor requirements, and sells little of its output to the local population, so high population density provides no attractions and considerable deterrence. Adjustment speed is faster in manufacturing than in total employment, which is surprising. A large percentage of black residents provides stronger attraction to manufacturing than to total employment. Not surprisingly, a dense interstate highway system is an even greater attraction to manufacturing than to total employment. We were surprised that high union membership is less of a deterrent to manufacturing than to total employment growth. Development bonds are a deterrent, at a low significance level, to manufacturing employment growth, their primary target.

Locational dummy coefficients are remarkably similar between the manufacturing and total employment growth regressions. Central cities attract manufacturing employment growth. Suburban and adjacent counties have similar attractions. A strong attraction of non-adjacent counties is apparent, which is consistent with the recent migration of manufacturing employment to counties distant from metropolitan areas. Regional dummies display mild deterrence of frostbelts manufacturing employment and stronger deterrence of locations to the Rocky Mountain region.

Surprisingly, R^2's are as high for regressions in Table 2 as for those in Table 1.

For most purposes, estimates of (7) and (8) are more interesting than those of (5) and (6). Table 3 shows coefficients of (7) and (8) calculated from estimates in Table 1, that is for population and total employment. Coefficients in Table 3 are estimates of long run, steady state effects of the S variables shown in the left column on the endogenous variables shown at the top of the second and third columns.

Not surprisingly, long run effects are large relative to short run effects. Most coefficients in Table 3 are ten or more times as large as corresponding coefficients in Table 1. This indicates that short run movements are misleading indicators of the magnitudes of long run effects. If a given exogenous variable has short run coefficients of opposite signs in the population and employment equations, even the direction of the long run effect may be the opposite from that of the short run effect.

Coefficients in Table 3 depend on units in which variables are measured. Table 4 presents elasticities of the two dependent variables with respect to exogenous variables that could be affected by government and private actions. The elasticities in Table 4 are long run effects calculated at sample means.

In both columns, elasticities with respect to family income are the largest. The calculations indicate that a one percent increase in median family income causes a long run

increase of 7 percent in the county's employment and a 5 percent increase in its population. These are strong effects indeed. As has been indicated, family income is correlated with educational attainment. Calculations not shown indicate similarly large elasticities of population and employment with respect to educational attainment.

The second largest elasticities in both columns are those with respect to the interstate highway density. Many counties have only one interstate highway. The calculated elasticities indicate that a second interstate highway which doubles the density might increase employment by 54 percent and populations by 17 percent. These are also large effects.

Table 3 Long Run and Reduced Form Equations for Total Employment and Population

	Total Employment	Population
Intercept	-0.442	-0.505
Percent Black	$0.249(10)^{-3}$	$-0.150(10)^{-3}$
Interstate Highway Density	0.195(10)	0.150(10)
Local Taxes per Capita	$-0.139(10)^{-4}$	$-0.168(10)^{-3}$
Crime	$-0.200(10)^{-6}$	$-0.243(10)^{-5}$
Family Income	$0.716(10)^{-4}$	$0.121(10)^{-3}$
Percent Union	$-0.127(10)^{-2}$	$-0.536(10)^{-3}$
Development Bonds	$0.651(10)^{-5}$	$0.276(10)^{-5}$
Central City Dummy	0.362	0.351
Adjacent Dummy	$-0.662(10)^{-3}$	-0.137
Non-adjacent Dummy	$0.364(10)^{-1}$	$-0.849(10)^{-1}$
New England	$-0.548(10)^{-1}$	0.142
Mid-Atlantic	-0.173	$-0.775(10)^{-1}$
Eastern Plains	-0.122	-0.275
Western Plains	$-0.839(10)^{-1}$	-0.222
East South Central	$0.325(10)^{-1}$	$0.143(10)^{-1}$
West South Central	$-0.967(10)^{-2}$	$0.543(10)^{-1}$
Rocky Mountain	-0.109	-0.261
Pacific	$0.812(10)^{-1}$	$-0.831(10)^{-1}$

The elasticities of population with respect to local taxes and of employment with respect to union membership are also large enough to suggest relevance for government policies. However, the estimated coefficients on which these elasticities are based did not have high significance levels. Other elasticities are both small and based on coefficient estimates with low significance levels.

Table 4 Elasticity Estimates

	Total Employment	Population
Percent Black	$0.341(10)^{-1}$	$-0.730(10)^{-2}$
Interstate Highway Density	0.537	0.174
Local Taxes per Capita	$-0.300(10)^{-1}$	-0.154
Crime	$-0.601(10)^{-2}$	$-0.307(10)^{-1}$
Family Income	0.681(10)	0.482(10)
Percent Union	-0.375	$-0.666(10)^{-1}$
Development Bonds	$0.356(10)^{-1}$	$0.643(10)^{-2}$

5. CONCLUSIONS

The calculations reported in this paper provide some insights and leave some mysteries.

First, the regressions indicate that population and employment are highly interactive, with each influencing the other strongly. Second, the regressions provide support for the notion that sunbelt amenities have had some effect in attracting people from the frostbelt. Generally, the regressions indicate little regional effect on employment.

Surprisingly, the regressions indicate that unmeasured characteristics make central cities attractive to residents and to employment. Residents, but not employment levels, are deterred by nonmetropolitan locations. The regressions indicate that residents are deterred by high local taxes, high crime rates and the presence of a large percentage of black residents. However, none of these effects is large or highly significant. High family income is a powerful attraction to both population and employment, and the low incomes typical in central cities must have been major causes of the migration of people and employment from them to other places. The analysis also suggests that the interstate highway system has had a significant effect on the location of both residents and jobs. Development bonds, the other direct government policy variable included in the regressions, appears to have had almost no effect in stimulating employment.

An overall impression left from the calculations is that migration of both population and employment from central cities to suburbs and, to a lesser extent, out of metropolitan areas is still something of a mystery. These movements are not accounted for by locational dummies, yet the effect of crime, taxes and racial mix appears to be relatively slight. Family income and the closely correlated educational attainment of the population have certainly been important, but may be as much consequences as causes of the overall process that has taken place.

ACKNOWLEDGEMENTS

The research and writing of this paper were carried out at the Federal Reserve Bank of Philadelphia. We are indebted to Barbara Lipman, Mark Siegel and Linda Heckert for computing and other assistance with the research. The views expressed herein are those of the authors and not those of the Federal Reserve Bank of Philadelphia.

REFERENCES

Carlino, G., 1985, "Declining City Productivity and the Growth of Rural Regions", *Journal of Urban Economics*, vol. 18, pp. 11-27.

Mills, E.S., 1986, "Metropolitan Central City Population and Employment Growth During the 1970's" in M.H. Peston and R.E. Quandt, (eds.), *Prices, Competition and Equilibrium*, Philip Allan Publishers Limited,Oxford, pp. 268-288.

Advances in Spatial Theory and Dynamics
Å.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 14

Some Theoretical Aspects of Spatial Equilibria with Public Goods

Å.E. Andersson, K. Kobayashi

1. INTRODUCTION

In some recent papers Andersson and Karlqvist (1976), Andersson and Ferraro (1983) and Ferraro (1984) have proposed and tested for Stockholm a simple model of accessibility and density distributions in a metropolitan area. The general idea behind the model, which departs a little from classic new urban economics, is that the city is essentially viewed as a set of public goods. The feature of public goods is due substantially to spatial interdependencies between decision makers.

Every locational decision is determined by the spatial allocation of public goods. If a person enters some small spatial unit he automatically becomes an input for agents besides himself. In this way, it comes close to the basic assumption of game theory, which explicitly assumes that the decision of a given decision maker is contingent upon simultaneous decisions by the other decision makers. In order to transform the Andersson-Ferraro framework into a more general one, some concepts of spatial gaming must be introduced, preferably with regard to the network and the technology of transportation. The next step in this analysis is thus the explicit introduction of some concepts from spatial gaming.

We can imagine a situation where all local public goods have already been allocated to nodes (zones). Every decision maker is assumed to be free to search for an optimal location, and will go for the most accessible node. Doing so, the agent will observe that the others do the same and thus recognizes that the second-best accessible node actually becomes advantageous, because of lower density, or in other words a higher local standard. There will, of course, sooner or later be an overflow of people in the node of secondary accessibility and thus a further spread of the population until nobody can enjoy higher utility by moving to a new location. In such a situation, a spatial equilibrium of population has been achieved. As can be induced from this description of a possible pattern of reactions of agents, free to search for their optimal location in an uncoordinated way, the criterion generating the equilibrium solution is that the level of individual utility should be the same in every location.

Our objectives in this paper are three-fold: (1) to study the spatial equilibrium of a population distribution, (2) to provide an alternative interpretation of the basic model of Andersson and Ferraro by viewing it as a mathematical programming model, (3) to give a more general framework, based on random utility theory, in which the basic model can be explained as a special case of a general spatial equilibrium model.

2. A RANDOM UTILITY MODEL WITH ACCESSIBILITY AND DENSITY AUGMENTATION

The basic feature of the Karlqvist-Andersson-Ferraro model is a principle of interdependencies between decision makers, which assumes that the decision of a given decision maker is contingent upon simultaneous decisions by other decision makers. Externalities will, in this context, be looked upon as not necessarily from distinct but rather as a subclass of public goods phenomena. Private goods are the goods such that the usage of them by one agent implies a proportional reduction in consumption by some other agents. Goods not fulfilling this property are called public goods. It should be noted that this definition of public goods is oriented to the commodity as an input to a household, a firm, an organization or any other economical unit. In this definition, we assume that the input of the commodity would have a noticable influence on the agent receiving it.

To simplify matters to the extreme, our study concerns a situation in which a fixed number of people is to be located on a transportation network with predetermined characteristics of nodes (zones) of the area. Let us assume that a utility function of an individual in the area is described in an additively separable form consisting of two parts: (1) one utility function common to the society and (2) one specific to the individual. The common utility function is assumed to include the accessibility and density measures as discussed in the Andersson-Ferraro model. Subscripts i (i=1,...,n) indicate nodes in the finite set of nodes. A general utility function is thus:

$$U_i^s = U_i(\alpha_i^y, S_i^z; \bar{z}_1,...,\bar{z}_i,...,\bar{z}_n, \bar{T}) + \bar{u}_i^s, \tag{1}$$

where

U_i^s = utility of individuals in node i,

α_i^y = accessibility to public good y from node i,

S_i^z = local standard of the node i with respect to characteristic z,

$\bar{z}_i$ = predetermined characteristics of the node i,

$\bar{T}$ = predetermined transportation system, and

$\bar{u}_i^s$ = deviation of utility of individual s from the common utility.

A description of interdependencies in a spatial public good analysis is most conveniently handled with an accessibility representation. Accessibility measures can be looked upon as the spatial counterpart of discounting. It thus represents the distribution of public goods in a simple way that imposes a very clear structure on the relation between people and their environment. In general, our accessibility measure has the following form:

$$\alpha_i^y = q_i(y_1,...,y_i,...y_n; \bar{T}); \ (i=1,...,n) \tag{2}$$

where y_i = the amount of public goods y in the node i. y_i could be $f(\bar{z}_i)$ or $g(X_i)$, where X_i is the population of zone i, and f and g are monotonously nondecreasing with $\bar{z}_i$ and X_i, respectively. For the simplicity of the mathematical discussion, let us assume that y takes the form $f(\bar{z}_i)$.

The next step is to include the density of population in a node and its inverse, i.e. spatial standard. The density, which is also an input for all members of the node, increases with the increase of population in the node. Thus the amount of the public good "availability of space" automatically decreases with increases of the population. Let us now introduce a spatial standard measure S_i^z which is an inverse counterpart measure of density and decreases with the increase of population.

$$S_i^z = h_i (X_i , \bar{z}_i); \ (i=1,...,n) \tag{3}$$

where h_i is 2-differentiable with X_i and $\bar{z}_i$, and monotonously increasing with X_i and nondecreasing with $\bar{z}_i$. It seems to be a reasonable hypothesis that the increase of density with respect to local public goods will decrease the common utility of an individual in the node. Thus the utility of an individual can be assumed to be monotonously increasing with an increase of the local standard. If we only consider accessibility to public goods but not to public "bads" it is furthermore reasonable that the increase of accessibility will be accompanied by an increase of the utility of individuals.

We can now fill the utility function (1) with a structural content. We have already included the positive benefits of agglomeration of individuals within the node by the accessibility measure, and the negative effect of congestion with individuals by the inverse measure of population, the local standard measure, in the common utility function. The measure of accessibility to local public goods can be represented as a function of predetermined transportation system $\bar{T}$ and local public goods of nodes $\bar{z}_i$ (i=1,...,n). In the same way, we can introduce a local standard measure in the common utility function, which should be looked upon as a function of population and local public goods of the node. In this line, a general form of utility function (1) can be rewritten to take the form

$$U_i^S = U_i (q_i (\bar{z}_1,...,\bar{z}_i,...,\bar{z}_n; \bar{T}), h_i (X_i, \bar{z}_i)) + \bar{u}_i^S. \tag{4}$$

A brief explanation must be attached to the specification of the utility function. Such a specification as given by expression (4) reflects two spatial features. The first is that the utility function of an individual in the society can be written in an additively separable form of two parts: the utility function common to the society and those specific to an individual. The function U_i in equation (4) reflects the representatives taste of the society and its value depends on the levels of accessibility and the local standard of the node. It is assumed that a common utility function U_i is differentiable and $\partial U_i/\partial q_i \geq 0$, $\partial U_i/\partial h_i \geq 0$, $\partial U_i/\partial \bar{X}_i \leq 0$ and $\partial U_i/\partial \bar{z}_i \geq 0$. On the other hand, $\bar{u}_i^S$ is a scalar and represents deviation of taste of an individual i from common utility. The second feature is that the stage for the location game is set by the predetermined characteristics $\bar{z}_i$ (i=1,...,n) and transportation system $\bar{T}$, which is the regulator of interactions. If the values of these exogenous variables, $\bar{z}_i$ and $\bar{T}$ are fixed, the common utility function U_i can be simplified so as to be a function of X_i, only.

3. THE SPATIAL CHOICE MODEL FOR POPULATION DISTRIBUTION

We now represent a spatial distribution model which describes the choice behavior of an individual choosing a node, given local public goods of nodes and the transportation system. The most commonly used spatial choice model to such individual behavior asserts that the conditional probability that a potential node is chosen by an individual varies directly with the measure of attractiveness of nodes. More recently, economists have also widely used variants of the "random utility model" (McFadden, 1975). In the random utility model it is assumed that a measure of attractiveness, attached to each alternative, completely determines the choice process of an individual.

It is important to note that the spatial choice model with random utility is an equilibrium model of spatial distribution in the sense that it describes an equilibrium state given the current values of the exogenous variables. In our demand situation of population distribution, some of the attributes of attractiveness of the nodes, i.e., the local standard measure of a node, depend upon the number of individuals in the same node, which is influenced by the relative attractiveness of other aspects of the nodes. If some of the parameters of the system affecting the local standard change, it is likely that the spatial equilibrium will also change. The resulting change in population distribution determines a new level of local standard of each node, which in turn affects its relative attractiveness. The spatial choice model can predict the new equilibrium situation that results from such a change in the following way: if a population distribution is in disequilibrium as a result of changes in the exogenous variables, each individual modifies his/her probability to select the node according to the random utility model until a new equilibrium state is reached at which the levels of local standard and spatial distribution of population equilibrate.

In this section we describe the spatial choice model briefly. In the next section we formally prove the existence of the equilibrium for population distribution, and demonstrate that under fairly general conditions the equilibrium is unique.

The common random utility model asserts that P_i, i.e. the probability that an individual will select node i, is given by the probability that the actual utility is maximized at node i,

$$P_i = \text{Prob}\{U_i^s \geq U_j^s \ ; \text{ for all } j=1,...,n\} \tag{5}$$

with

$$U_i^s = U_i + \tilde{u}_i^s \ ; \ (i=1,...,n) \tag{6}$$

where U_i^s and U_j^s represent the utility of an individual s obtained by selecting nodes i and j, respectively. Different assumptions on the distribution of an individual deviation of the utility $\tilde{u}_i^s$ result in different types of spatial choice models. It is well known that the multinomial logit model,

$$P_i = \exp(\alpha U_i) \Big/ \sum_{j=1}^{n} \exp(\alpha U_j), \tag{7}$$

is generated if the random component $\tilde{u}_i^s$ varies according to an extreme value (Gumbell, Weibull or Gnedenko) distribution.

$$G(\bar{u}_i^S,\alpha) = \alpha \exp[-\exp(-\alpha\bar{u}_i^S)] \tag{8}$$

which has standard deviation σ^2, given by

$$\sigma^2 = \pi^2/6\alpha^2. \tag{9}$$

The total population in the area is subdivided into node subsets. These subsets of the population are called X_i. X_i is given by

$$X_i = \bar{X} \exp(\alpha U_i)/ \sum_{j=1}^{n} \exp(\alpha U_j);(i=1,...,n) \tag{10}$$

where $\bar{X}$ is the total population of the area.

Now we move to the discussion concerned with the measurement of individual utilities. By putting

$$\bar{U}^S = \max_i \{U_i + \bar{u}_i^S\}, \tag{11}$$

since $\bar{u}_i^S$ is distributed according to a Gumbell function, we see that $\bar{U}^S$ also has a Gumbell function whose mean value is represented by:

$$E(\bar{U}^S) = \frac{1}{\alpha} \mathit{ln} \sum_{i=1}^{n} \exp(\alpha U_i) + \gamma, \tag{12}$$

where γ is the Euler constant. Expression (12) represents the expected value of an individual net utility, when the individual is randomly drawn from the population and provided that he will choose the optimal node that brings him the largest net utility. Any change in the predetermined public goods of nodes and the transportation system, $\bar{T}$, will directly affect the levels of $\bar{U}^S$.

4. THE EXISTENCE AND UNIQUENESS OF SPATIAL EQUILIBRIUM

In this section, we prove that the spatial distribution of the population has a unique equilibrium. Let us first consider an onto mapping F defined on an n-dimensional closed convex set D by the vector $F(X)=(F_1(X),..., F_n(X))$, whose components are given by

$$F_i(X) = \bar{X} \exp(\alpha U_i(X_i)) / \sum_{j=1}^{n} \exp(\alpha U_j(X_j)). \tag{13}$$

Let D be a compact convex set of R^n,

$$D = \{X: \sum_{i=1}^{n} X_i = \bar{X},\ X_i \geq 0 \text{ for all } i=1,...,n\}. \tag{14}$$

Clearly, if the spatial distribution of population has an equilibrium with the spatial choice model, then there exists a fixed point X^* such that

$$F(X^*) = X^*. \tag{15}$$

Theorem 1 There exists a fixed point for equation (15).

Proof: It is obvious that F is mapping from D onto itself. We can observe that F is continuous in X, since U_i is assumed to be continuous in X_i. Here let us introduce a function ϕ defined on DxD by

$$\phi(X,Y) = \langle F(X) - X,\ Y - X \rangle . \tag{16}$$

This function clearly satisfies the following assumptions of the Ky Fan inequality: (1) $\forall\ Y \in D,\ X \rightarrow \phi(X,Y)$ is lower semi-continuous, (2) $\forall\ X \in D,\ Y \rightarrow \phi\ (X,Y)$ is concave. The Ky Fan Theorem (Aubin and Ekeland, 1984, Chapter 6) shows that there exists $X^* \in D$ such that $\sup_{Y \in D} \phi(X^*, Y) \leq \sup_{Y \in D} \phi(Y, Y)$. That is, there exists $X^* \in D$ such that

$$\sup_{Y \in D} \langle F(X^*) - X^*,\ Y - X^* \rangle \leq \sup_{Y \in D} \langle F(Y) - Y,\ Y - Y \rangle = 0. \tag{17}$$

By taking $Y = F(X^*) \in D$, we get $(F(X^*)-X^*)^2 \leq 0$, that is X^*, is a fixed point of F.

We make use of a property of any Hicksian matrix in order to consider uniqueness of the equilibrium. We, in particular, show that the Jacobian of [F(X)-X] is Hicksian and hence the fixed point of F would be unique. Let us define the possible

$$\tilde{f}_i(X) = F_i(X) - X_i,\ (i=1,...,n). \tag{18}$$

We now let the Jacobian of $\tilde{f}(X) = (\tilde{f}_1(X),...,\ \tilde{f}_n(X))$ be $\tilde{F}(X) = \{\tilde{f}_{ij}(X)\}$ for (i, j = 1,...,n) so that

$$\tilde{f}_{ij}(X) = \partial \tilde{f}_i(X) / \partial X_j. \tag{19}$$

Lemma 1 Let $\tilde{F} = \{\tilde{f}_{ij}\}$ be an n·n matrix satisfying $\tilde{f}_{ij} \geq 0$ for all $i \neq j$. Then $\tilde{F}$ is Hicksian if and only if there exists a $V = \{v_1,..., v_n\} \geq 0$ satisfying $\tilde{F}'V' < 0$.

The proof of this Lemma can be found in Takayama (1974, Chapter 4).

Lemma 2 The matrix $\tilde{F}(X) = \{\tilde{f}_{ij}(X)\}$ is Hicksian for all $X \geq 0$.

Proof:

$$\tilde{f}_{ij}(X) = \begin{cases} -\alpha\bar{X} P_i P_j \dfrac{\partial U_j}{\partial X_j}, & \text{for } i \neq j; \\ \alpha\bar{X} P_i (1-P_i) \dfrac{\partial U_i}{\partial X_i} - 1, & \text{for } i = j; \end{cases} \tag{20}$$

where $P_i = \exp(\alpha U_i(X_i)) / \sum_j \exp(\alpha U_j(X_j)) \geq 0$, (i=1,...,n). Since we assume $\partial U_i(X_i)/\partial X_i \leq 0$, F(X) satisfies $f_{ij}(X) \geq 0$ for all $i \neq j$. Let V = (1,...,1) and define $\Gamma(X)$ = $\{\gamma_1(X),..., \gamma_n(X)\}$ where $\Gamma(X) = \tilde{F}(X)'V'$. Then

$$\begin{aligned} \gamma_j(X) &= \sum_{i=1}^{n} \tilde{f}_{ij}(X) \\ &= \alpha\bar{X} \left\{ \sum_{i\neq j} P_i P_j \left(\frac{\partial U_j}{\partial X_j} - \frac{\partial U_i}{\partial X_j}\right) \right\} - 1 \\ &= -1. \end{aligned} \tag{21}$$

Thus $\gamma_j(X) < 0$ for all $X \geq 0$ and $\alpha > 0$. Hence, by Lemma 1, we understand that F(X) is Hicksian for all $X \geq 0$.

Theorem 2 There exists a unique fixed point for equation (15) for arbitrary $\alpha > 0$.

Proof: Let us define a rectangular region G in R^n by

$$G = \{(X_1,..., X_n): 0 \leq X_i \leq \bar{X};\ (i=1,...,n)\}. \tag{22}$$

Consider the mapping $\tilde{f}(X)$ from G. If $\tilde{F}(X)$ is Hicksian for all $X \in G$, then the mapping $\tilde{f}(X)$ is one to one for all $X \in G$ (Nikaido, 1968, Chapter 7). Needless to say, an equilibrium point X* satisfies $\tilde{f}(X^*) = 0$ and $X^* \in D$. Since D is obviously a subset of G, $\tilde{f}(X)$ must also be one to one in D. Hence, there is only one $X^* \in D$ satisfying $\tilde{f}(X^*) = 0$.

5. A MATHEMATICAL PROGRAMMING FORMULATION OF THE RANDOM UTILITY MODEL

Before investigating the property of a fixed point of equation (15), it is important to consider an alternative interpretation of the random utility model by reformulating it as a mathematical programming model. Denote the size of the population in the area as a whole by $\bar{X}$. Then total expected utility T(U) is given by

$$T(U) = \frac{\bar{X}}{\alpha} \ell n \sum_{i=1}^{n} \exp(\alpha U_i). \tag{23}$$

Here, we consider T(U) as a function of the utility levels of nodes in the area. By making a conjugate function of (23), we can construct a mathematical programming model.

Lemma 3 For arbitrary $\alpha > 0$, the function $T(U)$: $R^n \rightarrow R$ is proper convex, and its conjugate function $T^*(X)$: $R^n \rightarrow [-\infty,+\infty]$ is given by

$$T^*(X) = \begin{cases} \frac{1}{\alpha} \sum_{i=1}^{n} X_i \ell n(X_i/\bar{X}) & ; \text{if } \sum_{i=1}^{n} X_i = \bar{X},\ X_i \geq 0;\ (i=1,...,n); \\ +\infty & ; \text{otherwise} \end{cases} \tag{24}$$

Note that $0 \cdot \log 0 = 0$.

Proof: For $\bar{X} = 0$, lemma 3 holds trivially. Let us suppose that $\bar{X} \neq 0$. Clearly, T(U) is a proper convex function. In general, a conjugate function $f^*(X)$ to a certain proper convex function f(U) is defined as follows:

$$f^*(X) = \sup_{U \in S} \{ < U,X > - f(U) \} \quad \text{for all } X \in \hat{S};$$

where S is the region, where a function f is defined, and

$$\hat{S} = \{X : \sup_{U \in S} [< U,X > - f(U)] < \infty \} .$$

Let us now consider the conjugate function $T^*(X)$ to the total expected utility function (23). Since the function T(U) is proper convex, the value of $T^*(\bar{X})$ is equivalent to the optimal value of the following convex programming problem:

$$\text{minimize } T^{**}(U) = \sum_{i=1}^{n} X_i \{U_i - \frac{1}{\alpha} \ell n(X_i/\bar{X})\};$$

$$\text{subject to } \sum_{i=1}^{n} X_i = \bar{X};\quad X_i \geq 0 ;\quad (i=1,...,n) . \tag{25}$$

If $X_i > 0$, i = 1,...,n, the Lagrange function of (25) is

$$L(X,\lambda) = \sum_{i=1}^{n} X_i \{U_i - \frac{1}{\alpha} \ell n(X_i/\bar{X})\} + \lambda(\bar{X} - \sum_{i=1}^{n} X_i) .$$

Let us consider the dual problem of (25). The optimal value function of the dual problem (25), $G(\lambda)$: $R \rightarrow R$ is

$$G(\lambda) = \max_{X} \{ L(X,\lambda): X \in R^n \}$$

$$= \lambda\bar{X} + \sum_{i=1}^{n} \frac{\bar{X}}{\alpha} \exp(\alpha U_i - \alpha\lambda - 1) . \tag{26}$$

From the duality theorem of the convex programming problem, we get

$$T^{**}(U) = \min_{\lambda} \{G(\lambda): \lambda \in R\}$$

$$= \frac{\bar{X}}{\alpha} \ell n \sum_{i=1}^{n} \exp(\alpha U_i) . \tag{27}$$

Thus the bi-conjugate function $T^{**}(U)$ is equivalent to $T(U)$, since the optimal solution of (25) is

$$X_i = \bar{X} \exp(\alpha U_i) / \sum_{k=1}^{n} \exp(\alpha U_k) > 0,$$

the assumption, $X_i > 0$, $i = 1,...,n$, is automatically satisfied.

From Lemma 3, we can now get an equivalent mathematical reformulation of the total expected utility function (23). Hence we get the following proposition.

Proposition 1 The optimal value function of the problem P-1 is equivalent to the total expected utility measure (23).

$$\text{(P-1)} \quad T(U) = \max_{X} \left\{ -\frac{1}{\alpha} \sum_{i=1}^{n} X_i \ell n (X_i/\bar{X}) + \sum_{i=1}^{n} U_i X_i \right\};$$

subject to

$$\sum_{i=1}^{n} X_i = \bar{X} ; X_i \geq 0; (i=1,...,n) . \tag{28}$$

6. ECONOMIC INTERPRETATION OF THE SPATIAL EQUILIBRIUM

In this section, we prove an equivalent formulation of the spatial equilibrium of population distribution with local public goods, and provide an alternative interpretation of the basic model of Andersson and Ferraro by viewing it as a mathematical programming model. We now inquire about the kind of program equation (15) has as the first-order necessary conditions for its optimal solution.

Proposition 2 The first-order necessary conditions for the optimality of the problem P-2:

$$\text{(P-2)} \quad \max_{X} \{-\frac{1}{\alpha} \sum_{i=1}^{n} X_i \ell n\,(X_i/\bar{X}) + \sum_{i=1}^{n} \int_0^{X_i} U_i(X_i)dX_i\};$$

$$\text{subject to} \quad \sum_{i=1}^{n} X_i = \bar{X}\ ;\ X_i \geq 0;\ (i=1,...,n) \tag{29}$$

are equivalent to equation (15).

Proof: The optimal conditions of the problem P-2 can be given by

$$-\frac{1}{\alpha}\,(\ell n\,(X_i/\bar{X})+1) + U_i(X_i) - \lambda = 0;\ (i=1,...,n), \tag{30}$$

where λ is a Lagrange multiplier. By eliminating λ from (30), we can see that the optimal solution of the problem P-2 satisfies

$$X_i = \frac{\bar{X}\,\exp(\alpha U_i\,(X_i))}{\sum_{k=1}^{n} \exp(\alpha U_k\,(X_k))}\ . \tag{31}$$

It must be emphasized that parameter α in (31) is related to the standard deviation of the probabilistic distribution representing the dispersion of utility over individuals. It is interesting to briefly explore the $\alpha \to \infty$ limit, which implies a situation where every individual has the same utility function. It is easy to use our formulation of the problem to show that the maximization of the second term within our overall objective function should then coincide with that given by the Andersson-Ferraro model. In fact, when $\alpha \to \infty$, the first term disappears from the objective function. In the $\alpha \to \infty$ limit, the optimal condition can be given by

$$U_1(X_1) = U_2(X_2) = ... = U_n(X_n), \tag{32}$$

which is equivalent to the equilibrium condition of the Andersson-Ferraro model. The objective function of (29) is a linear combination of an entropy function and a description of a spatial gaming situation, presumed in the Andersson-Ferraro model. The parameter α describes the trade-off relation between a spatial randomization and a spatial gaming situation. The value of α would typically be inferred from a calibration of the spatial choice model.

Let us now investigate the uniqueness of the optimal solution of the problem P-2. It is obvious from Proposition 2 that the problem P-2 has a unique global optimal solution. Then this problem gives a unique fixed point of (15).

Proposition 3 There exists a global optimal solution of problems P-2 for arbitrary $\alpha > 0$.

Proof: The sufficient condition for the global minimum of the problem P-2 is that the Hessian matrix of second derivatives of the objective function (29) is negative definite for all $X \in D$. Each element of the Hessian matrix can be directly calculated from (29), v.z.,

$$h_{ij} = \begin{cases} 0 & ; \text{ if } i \neq j; \\ -\dfrac{1}{\alpha X_i} + \dfrac{\partial U_i(X_i)}{\partial X_i} & ; \text{ if } i = j. \end{cases} \tag{33}$$

Since we assume $\partial U_i(X_i)/\partial X_i \leq 0$, h_{ii} for all $X_i \in D$. From Proposition 2 and 3, we get the following theorem.

Theorem 3 The mathematical programming problem P-2 gives a fixed point of equation 15.

7. THE SPATIAL EQUILIBRIUM AND DUALITY

Let us first consider the Lagrange function of (29), $L: R^n \times R \rightarrow [-\infty, +\infty]$:

$$L(X,\lambda) = \begin{cases} f(X) + \lambda(\bar{X} - \sum_{i=1}^{n} X_i) ; & \text{if } \lambda \geq 0; \\ +\infty \quad ; & \text{if } \lambda < 0; \end{cases} \tag{34}$$

where $f(x) = -\dfrac{1}{\alpha} \sum_{i=1}^{n} X_i \ell n(X_i/\bar{X}) + \sum_{i=1}^{n} \int_0^{X_i} U_i(X_i)dX_i$, and introduce two functions,

$\theta: R^n \rightarrow [-\infty, +\infty]$ and $\omega: R \rightarrow [-\infty, +\infty]$ defined by

$$\theta(X) = \inf_{\lambda \in R} L(X,\lambda), \tag{35}$$

and

$$\omega(\lambda) = \sup_{X \in R^n} L(X,\lambda) . \tag{36}$$

Then the problem (29) is equivalent to the problem P-3:

$$\text{(P-3)} \quad \text{maximize } \theta(X), \tag{37}$$

and the Lagrangean dual problem of (37) is

$$\text{(D-3)} \quad \text{minimize } \omega(\lambda) . \tag{38}$$

Let us here define the constraint mapping, $S: R \rightarrow R(R^n)$, and the optimal value function, $\phi(u)$: $R \rightarrow [-\infty, +\infty]$ of (29) by

$$S(u) = \{X \in R^n: \sum_{i=1}^{n} X_i = \bar{X} + u,\ X_i \geq 0\}, \tag{39}$$

and

$$\phi(u) = \sup_{X \in R^n} \{f(X): X \in S(u)\} \tag{40}$$

where $u \in R$ is a parameter. $\phi(u)$ is a nondecreasing and proper concave function, since $f(x)$ and $S(u)$ are proper concave (Rockafellar, 1970). Moreover, we define the function $H: R^{n+1} \to [-\infty, +\infty]$ by

$$H(X,u) = f(X) + \delta_{S(u)}(X) \tag{41}$$

where $\delta_{S(u)}(X): R \to [-\infty, 0]$ is the indicator function of the constraint mapping $S(u)$:

$$\delta_{S(u)}(X) = \begin{cases} 0 & ;\ \text{if } X \in S(u); \\ -\infty & ;\ \text{if } X \notin S(u) \end{cases}$$

$H(X,u)$ is closed concave since $f(X)$ and $S(u)$ are concave. Then we get

$$\phi(u) = \sup_{X \in R^n} H(X,u)\ . \tag{42}$$

From equations (34), (41) and (42), we get the following lemmas:

Lemma 4 There exist the following relations between the Lagrange function (34), $L: R^{n+1} \to [-\infty, +\infty]$, and the function (41), $H: R^{n+1} \to [-\infty, +\infty]$:

$$L(X,\lambda) = \sup_{u \in R} \{H(X,u) - < \lambda, u > \}; \tag{43}$$

$$H(X,u) = \inf_{\lambda \in R} \{L(X,\lambda) + < \lambda, u > \}. \tag{44}$$

Proof: Trivial from definitions of (34) and (41).

Lemma 5 For arbitrary $\lambda \in R$, we get the relation

$$\omega(\lambda) = -\phi^*(\lambda), \tag{45}$$

where $\phi^*(\cdot)$ is the conjugate function of the objective value function $\phi(u)$.

Proof: By Lemma 4 and equation (41), we get

$$\omega(\lambda) = \sup_{X \in R^n} L(X,\lambda)$$

$$= \sup_{X \in R^n} \sup_{u \in R} \{H(X,u) - <\lambda,u>\}$$

$$= \sup_{u \in R} \{\phi(u) - <\lambda,u>\}$$

$$= - \phi^*(\lambda).$$

Lemma 6 For the dual problem (D-3),

$$\inf_{\lambda \in R} \omega(\lambda) = \phi^{**}(0). \tag{46}$$

Proof: From Lemma 5, we have

$$\inf_{\lambda \in R} \omega(\lambda) = \inf_{\lambda} \{<\lambda, 0> - \phi(\lambda)\}$$

$$= \phi^{**}(0)$$

where $\phi^{**}(0)$ is the bi-conjugate function of $\phi(u)$ evaluated at $u = 0$. Since the optimal value function $\phi(u)$ is proper concave,

$$\phi^{**}(0) = \phi(0).$$

From the above Lemma, we get the duality theorem.

Theorem 4 For the primal problem (P-3) and the dual problem (D-3),

$$\sup_{X \in R^n} \theta(X) = \inf_{\lambda \in R} \omega(\lambda). \tag{47}$$

Proof: Obvious from Lemma 6.
Now let us consider dual problem D-3. From Lemma 5, the dual problem D-3

$$\inf_{\lambda \in R} \omega(\lambda) = \inf_{\lambda \in R} \{\phi^*(\lambda)\}$$

$$= \inf_{\lambda \in R} \sup_{u \in R} \{\phi(u) - <\lambda,u>\}$$

$$= \inf_{\lambda \in R} \sup_{X \in R^n} \{f(X) - \lambda(\sum_{i=1}^{n} X_i - \bar{X})\}$$

$$= \inf_{\lambda \in R} \{ \lambda \bar{X} + \Psi(X,\lambda)\}, \tag{48}$$

where $\Psi(X,\lambda) = \sup_{X \in R^n} \{f(X) - \lambda \sum_{i=1}^{n} X_i \}$. From the definition of f(X)

$$\Psi(X,\lambda) = \max_{X} \{-\lambda \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \frac{1}{\alpha} X_i \ell n(X_i/\bar{X}) + \sum_{i=1}^{n} \int_0^{X_i} U_i(X_i)dX_i\}$$

$$= \max_{X} \{ \sum_{i=1}^{n} X_i(-\lambda - \frac{1}{\alpha}\ell n(X_i/\bar{X}) + U_i) - \sum_{i=1}^{n} \int_{U_i(0)}^{U_i(\lambda,X)} g_i(u_i)du_i \} \quad (49)$$

which is obtained by integrating the third term by parts and where $g_i(u_i)$ is the inverse function of the utility function, which exists since the utility function is assumed to be monotone and nondecreasing. From the first-order optimality condition of (49), we get,

$$-\lambda - \frac{1}{\alpha} \ell n(X_i/\bar{X}) + U_i - \frac{1}{\alpha} = 0. \quad (50)$$

By substituting (50) into (49), we get the dual problem:

$$\min_{\lambda} \{\lambda \bar{X} + \frac{1}{\alpha} \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \int_{U_i(0)}^{U_i(\lambda,X)} g_i(u_i)du_i \}. \quad (51)$$

The first-order optimality condition of the dual problem is

$$\bar{X} = \sum_{i=1}^{n} g_i(U_i(\lambda,X)) = \sum_{i=1}^{n} X_i . \quad (52)$$

From (50) and (52), we get

$$\lambda^* = \frac{1}{\alpha} \ell n \sum_{i=1}^{n} \exp(\alpha U(X_j^*)) - \frac{1}{\alpha} , \quad (53)$$

where λ^*, X_i^* are the optimal solutions of the dual problem. By substituting (53) to (51), the optimal value of the dual problem is

$$\omega(\lambda^*) = \frac{1}{\alpha} \ell n \sum_{i=1}^{n} \exp(\alpha U_i(X_i^*)) - \sum_{i=1}^{n} \int_{U_i(0)}^{U_i(X_i^*)} g_i(u_i)du_i$$

$$= \frac{1}{\alpha} \ell n \sum_{i=1}^{n} \exp(\alpha U_i(X_i^*)) + \sum_{i=1}^{n} (\bar{U}_i(X_i^*) - U_i(X_i^*)) X_i^*$$

where $\bar{U}_i(X_i^*) = \int_0^{X_i^*} U_i(X_i)dX_i/X_i^*$. (54)

From theorem 4, it is guaranteed that the value of $\omega(\lambda^*)$ is equal to the optimal value of $\theta(X^*)$. Equation (54) suggests that the optimal value of the problem P-3 can be interpreted as a linear combination of a social cost caused by the congestion in local public goods of the nodes and a total group expected utility evaluated at the equilibrium point X_i^*, i=1,...,n.

CONCLUSION

This paper develops an equilibrium theory for the analysis of accessibility and density distributions. We analyze the issue of congestion of nodes and impacts on the resulting spatial equilibrium of population. The key results of this paper are (1) to provide an alternative interpretation of the basic model of Andersson and Ferraro by viewing it as a mathematical programming model, and (2) to give a more general framework, based upon the random utility model, in which the basic model can be regarded as a special case of a general spatial equilibrium model.

Although still remote from a comprehensive theory of accessibility and density distributions, this paper describes some initial elements which may serve as a model spatial equilibrium model of a single type of agents in the design of such a theoretical framework. We believe that structural forms for the interrelationships between different types of agents can be constructed along with some mathematical programming formulation of the spatial equilibrium as discussed in this paper. Further research is still needed, however, with respect to the linking mechanism which describes the spatial interdependences between multi-type agents in the region into a unified analytical framework of general spatial equilibrium.

REFERENCES

Andersson, Å.E. and G.V.G. Ferraro, 1983, Accessibility and Density Distribution in Metropolitan Areas: Theory and empirical studies. *Papers of Regional Science*, 52:141-158.

Andersson, Å.E and A. Karlqvist, 1976, Population and Capital in Geographical Space. The Problem of General Equilibrium Allocation, in J. Los and W. Los (eds.), *Computing Equilibria: How and Why*, pp. 183-195, North-Holland, Amsterdam.

Aubin, J-P. and I. Ekeland, 1984, *Applied Nonlinear Analysis*, John Wiley & Sons, New York.

Ferraro, G., 1984, Localizzazioni Residenziali entro Aree Metropolitane: Comportamenti individuali ed assetti di equilibrio. *Richerche Economiche*, Vol. 3 38, NO.2.

McFadden, D., 1975, Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka, (ed.), *Frontiers in Econometrics*, Academic Press, New York.

Nikaido, H., 1968, *Convex Structure and Economic Theory*. Academic Press, New York.

Rockafellar, R.R., 1970, *Convex Analysis*, Princeton University Press, New Jersey.

Takayama, A., 1974, *Mathematical Economics*. The Dryden Press, Hinsdale, Ill.

Advances in Spatial Theory and Dynamics
Å.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 15

A General Dynamic Spatial Price Equilibrium Model with Gains and Losses

A. Nagurney

1. INTRODUCTION

The central issue in the study of dynamic spatial price equilibrium problems is the computation of equilibrium regional production and consumption, as well as, interregional and intertemporal commodity flow patterns. In a prescribed manner, the supplies and demands for the commodity depend on the prices, the inventory costs on the quantities stored, and the transportation costs on the shipment pattern. Applications of such models arise in a variety of fields such as agriculture, regional science, and energy.

Samuelson (1957) and Takayama and Judge (1964, 1971) introduced temporal spatial price equilibrium models and proposed algorithms for the computation of the equilibrium, but these models could not handle situations in which, for example, the transportation and inventory costs were no longer constant, but rather, functions of the entire interregional and intertemporal flow pattern. Moreover, the supply and demand functions were usually assumed to be linear and integrable. Inventorying at supply and demand markets was not distinguished in such models until more recently by Guise (1979), and the case of backordering was also ignored. Finally, there was no treatment of such phenomena as shrinkage/accretion of commodities over space and time due to pilferage, perishability, losses or, respectively, amplification in value or volume.

Takayama and Uri (1983) and Takayama et al. (1984) relaxed the symmetry assumption on the functions, but retained the linearity and formulated temporal models as linear complementarity problems. However, the computational experience, so important for the operationalism of the models, was limited. Thore (1986), on the other hand, introduced multipliers in a symmetric model and assumed that the inventory costs were fixed and identical between time periods.

Nagurney and Aronson (1988), in an effort to circumvent some of the afore-mentioned restrictive assumptions, proposed a dynamic spatial price equilibrium model over a finite time horizon and provided alternative variational inequality formulations of the equilibrium conditions. In the special case of only a single time period, this model is the well-known static spatial price equilibrium problem (s.p.e.p.) studied by Florian and Los (1982), Friesz, et al. (1983), Friesz et al. (1984), Pang (1984, 1985), Dafermos and Nagurney (1984, 1987), Jones et al. (1984a,b), and Nagurney (1987). Their formulation is based crucially on the visualization of the dynamic s.p.e.p. as a multiperiod network. The computational experience with Gauss-Seidel type serial linearization decomposition schemes by supply markets in time and by demand markets in time suggested that

problems with as many as twenty supply markets and twenty demand markets over ten time periods can be effectively solved in the order of minutes.

In Section 2 of this paper we introduce a general dynamic spatial price equilibrium model to handle gains and losses through the use of multipliers. This framework extends the dynamic model of Nagurney and Aronson (1988) and allows for a more realistic representation of, for example, agricultural markets for perishable commodities and financial markets with the characteristic credit multipliers. This model generalizes both the static and dynamic deterministic models given in Thore (1986). We then define the governing equilibrium conditions with the incorporation of multipliers and give alternative variational inequality formulations of the problem over Cartesian products of sets (see also, e.g., Nagurney (1987), Pang (1985)). The structures of the variational inequalities are very similar to those given in Nagurney and Aronson (1988).

In Section 3 we consider the solution of the general model and outline the Gauss-Seidel serial linearization decomposition method by demand markets in time for the case of gains and losses. In Section 4 we outline the adaptation of the equilibration scheme by demand markets in time to handle gains and losses, as well.

In Section 5 we present computational experience for the adaptation of the equilibration scheme by demand markets in time for randomly generated examples. In Section 6 we then provide computational experience for the Gauss-Seidel type decomposition method by demand markets in time for the solution of the general model with gains and losses for a variety of scenarios.

2. A GENERAL DYNAMIC SPATIAL PRICE EQUILIBRIUM MODEL WITH GAINS AND LOSSES

In this section we present a generalized version of the dynamic spatial price equilibrium model introduced by Nagurney and Aronson (1988). We consider a finite time horizon and partition the horizon into discrete time periods, t, t=1,...,T. We assume that a certain commodity is produced at m supply markets and is consumed at n demand markets. We denote a typical supply market by i and a typical demand market by j. We number the supply markets from 1 through m and the demand markets from m+1 through m+n.

Let s_{it} denote the quantity of the commodity produced at supply market i in the t-th period and let d_{jt} denote the demand associated with the demand market j at the t-th period. We arrange the supplies into T-tuples of vectors $\{s_1,...,s_T\}$ in R^m. Then we incorporate the above T-tuples into a single vector s in R^{mT}. Similarly, we arrange the demands into T-tuples of vectors $\{d_1,...,d_T\}$ in R^n. Then we incorporate the T-tuples into a single vector d in R^{nT}.

We now discuss the commodity shipments, the inventoried and backordered quantities, and the associated finite positive multipliers to allow for gains and losses.

Let x_{itjt} denote the amount of the commodity shipped from supply market i to demand market j in period t and let a_{itjt} denote the associated multiplier. Hence, if at the beginning of time period t an amount x_{itjt} is shipped from i to j, the amount that will arrive is given by $a_{itjt}\, x_{itjt}$. Let x_{itit+1} denote the amount of the commodity inventoried at supply market i from time period t to t+1 and let x_{jtjt+1} denote the amount inventoried at demand market j from t to t+1, where a_{itit+1} and a_{jtjt+1} denote the respective multipliers. The amounts of the commodity that remain in inventory at markets i and j at the beginning of time period t+1 are given, respectively by $a_{itit+1}\, x_{itit+1}$ and $a_{jtjt+1}\, x_{jtjt+1}$. Finally, we denote the amount backordered at demand market j from time period t to t-1 by x_{jtjt-1} and its corresponding multiplier by a_{jtjt-1}. We group the commodity shipments $\{x_{itjt}\}$ into a vector x_1 in R^{mnT}, the quantities inventoried at the supply

markets, $\{x_{itit+1}\}$, into a vector x_2 in $R^{m(T-1)}$, the quantities inventoried at the demand markets, $\{x_{jtjt+1}\}$, into a vector x_3 in $R^{n(T-1)}$, and the quantities backordered $\{x_{jtjt-1}\}$ into a vector x_4 in $R^{n(T-1)}$. We then group the vectors x_1,x_2,x_3,x_4 into a vector x in $R^{mnT+m(T-1)+2n(T-1)}$.

Similarly we group the shipment multipliers $\{a_{itjt}\}$ into a vector a_1 in R^{mnT}, the inventory multipliers $\{a_{itit+1}\}$ and $\{a_{jtjt+1}\}$ into vectors $\{a_2\}$ and $\{a_3\}$ in $R^{m(T-1)}$ and $R^{n(T-1)}$, respectively, and the backorder multipliers $\{a_{jtjt-1}\}$ into a vector a_4 in $R^{n(T-1)}$. We then group the vectors a_1, a_2, a_3, a_4 into a vector a in $R^{mnT+m(T-1)+2n(T-1)}$. Note that in most applications of interest a multiplier will lie in the range (0,2); in the case of perishability, pilferage, or loss in the range (0,1); in the case of credit multipliers as in financial applications typically in the range (1,2). Usually the multipliers associated with the backorder quantities will be set equal to one. In the case that all the multipliers are equal to one, the model is that of Nagurney and Aronson (1988).

We associate with each supply market i at each time period t a supply price π_{it} and with each demand market j at each time period t a demand price p_{jt}. We arrange the supply prices into T-tuples of vectors $\{\pi_1,...,\pi_T\}$ in R^m. Then, we incorporate this T-tuple into a vector π in R^{mT}. Similarly, we arrange the demand prices into T-tuples of vectors $\{p_1,...,p_T\}$ in R^n. Then we incorporate this T-tuple into a single vector p in R^{nT}.

We denote the transportation cost of the commodity from supply market i to demand market j at period t by c_{itjt}. We let c_{itit+1} denote the inventorying cost at supply market i from t to $t+1$, and we let c_{jtjt+1} denote the inventorying cost at demand market j from t to $t+1$. We denote the backordering cost at demand market j from t to $t-1$ by c_{jtjt-1}. We group the transportation costs $\{c_{itjt}\}$ into a vector c_1 in R^{mnT}, the supply market inventorying costs, $\{c_{itit+1}\}$ into a vector c_2 in $R^{m(T-1)}$, the demand market inventorying costs $\{c_{jtjt+1}\}$ into a vector c_3 in $R^{n(T-1)}$, and the backordering costs $\{c_{jtjt-1}\}$ into a vector c_4 in $R^{n(T-1)}$. We then group c_1, c_2,c_3, c_4 into a single vector c in $R^{mnT+m(T-1)+2n(T-1)}$. We assume that the transportation, inventorying, and backordering costs are nonnegative.

We now review the construction of the dynamic price equilibrium network of Nagurney and Aronson (1988). (For a graphical representation see Figure 1.) For each period t, $t=1,...,T$ we construct m supply market nodes, denoted by the pairs $it,...,mt$ representing the supply markets at period t, and n demand market nodes, denoted by the pairs $(m+1)t,...,(m+n)t$, representing the demand market at period t. For each time period t, we construct mn transportation links, a typical one originating at a node it and terminating at a node jt. We denote such a link by $itjt$. Hence, the total number of transportation links in G is mnT. From each supply market node it, we then construct a supply market inventory link, denoted by $itit+1$, terminating in supply market node $it+1$; and from each demand market node jt we construct a demand market inventory link, denoted by $jtjt+1$, terminating in demand market node $jt+1$. There are a total of $m(T-1)$ supply market inventory links and $n(T-1)$ demand market inventory links. From each demand market node jt, we further construct a demand market backorder link, denoted by $jtjt-1$, terminating in demand market node $jt-1$, yielding a total of $n(T-1)$ backorder links. The total number of links in G is therefore $mnT+m(T-1)+2n(T-1)$.

A sequence of links originating in supply market node it and terminating in demand market node jt' induces a path. We refer to a typical path from a supply market node to a demand market node by r. We consider only paths without cycles.

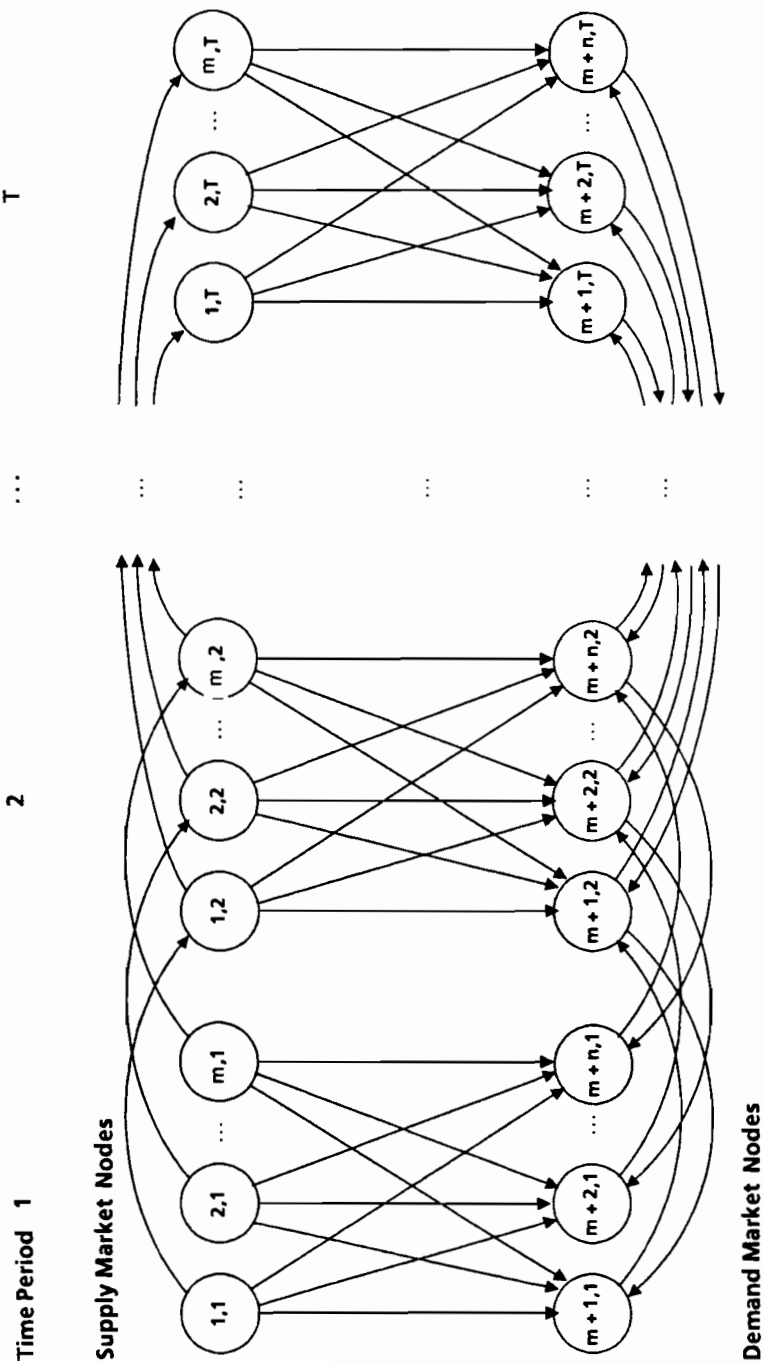


Figure 1 A Network Representation of the General Dynamic Spatial Price Equilibrium Problem

In this network representation, we now associate with each defined link itjt', an $x_{itjt'}$ and a finite, positive multiplier $a_{itjt'}$. We denote then the flow on a path r by x_r which represents the quantity of the commodity utilizing path r and the effective cost on path r by C_r. We let P denote the set of paths in the network, P^{it} the set of paths originating in supply market node it, $P_{jt'}$ the set of paths terminating in demand market node jt', and $P_{itjt'}$ the set of paths originating in supply market node it and terminating in demand market node jt'. Let np, npit, $np_{jt'}$, and $np_{itjt'}$ denote, respectively, the number of paths in the network, the number of paths originating in supply market it, the number of paths terminating in demand market node jt', and the number of paths originating in supply market node it and terminating in demand market node jt'.

We group the C_r's into a vector C in R^{nP} and the x_r's into a vector y in R^{nP}.

Note that in applications, inventorying may be allowed only at the supply markets or at the demand markets or at certain supply and certain demand markets in certain time periods. Moreover, in certain time periods, due for example, to seasonalities, certain supply or demand markets may not exist. For such reasons the number of links and the number of paths in the network representation will be application dependent and, we expect that in most cases will be of a dimension lower than that in the general graphical representation in Figure 1. For the same reasons, the dimensionality of the various vectors will also be application dependent.

In this model, unlike the one given in Nagurney and Aronson (1988), the case of restrictions on inventorying to allow for perishability and losses is no longer handled via constraints imposed a priori, but, rather, instrinsically through the use of multipliers and the equilibrium conditions.

Towards this end we define the chain and path multipliers, $a_{(itjt')r}$ and a_r, for path r, respectively by

$$a_{(itjt')r} = \prod_{\mathbf{itjt'} < itjt'} a_{\mathbf{itjt'}} \quad a_{itjt'}^{\delta_{(\mathbf{itjt'})r}, \delta_{(itjt')r}} \quad a_r = \prod_{itjt' \in r} a_{itjt'} \tag{1}$$

where $\mathbf{itjt'} < itjt'$ denotes a link $\mathbf{itjt'}$ which precedes link itjt', in a path containing itjt' and $\delta_{(itjt')r} = 1$ if link itjt' is contained in path r and 0 otherwise. If itjt' is the first link in a path we define $a_{(itjt')r}$ to be equal to one.

The quantities produced and consumed must satisfy the following conditions:

$$s_{it} = \sum_{r \in P_{it}} x_r \quad , \quad d_{jt'} = \sum_{r \in P_{jt'}} a_r x_r \tag{2}$$

$$x_{itjt'} = \sum_r a_{(itjt')r} x_r = \sum_r x_{r(itjt')} \tag{3}$$

where $x_{r(itjt')}$ represents the effective contribution of the flow on path r on the quantity $x_{(itjt')}$.

The effective cost on path r is given by

$$C_r = \sum_{itjt'} a_{(itjt')r} c_{itjt'}. \tag{4}$$

A spatial price equilibrium with gains and losses consisting of prices, shipments, and quantities inventoried and backordered is established if the value of the resultant quantity of the commodity arriving at a demand market does not exceed the price of the quantity produced at the supply market plus the effective cost of inventorying, shipping, backordering, etc., if there is to be a positive commodity flow between the pair of supply and demand markets.

Following Takayama and Judge (1971) and Nagurney and Aronson (1988), the temporal and spatial price equilibrium conditions here take the form: for all pairs of supply market nodes and demand market nodes it, jt', i=1,...,m; j=1,...,n; t=1,...,T; t'=1,...,T, and all paths r joining market nodes it, jt'

$$\pi_{it} + C_r \quad \begin{cases} = a_r p_{jt'}, & \text{if } a_r x_r > 0 \\ \geq a_r p_{jt'}, & \text{if } a_r x_r = 0. \end{cases} \tag{5}$$

In the special case where all of the multipliers are equal to one, equilibrium conditions (5) are those governing the dynamic model in Nagurney and Aronson (1988). In the special case where we have a single time period t, equilibrium conditions (5) are those given in Thore (1986). If we then further assume that all of the $a_{itjt'}$ are equal to one, then equilibrium conditions (5) reduce to the static spatial price equilibrium conditions, in which C_r consists of only the transportation cost (see, e.g. Dafermos and Nagurney (1985), Nagurney (1987)).

We shall now discuss the supply, demand, transportation, inventorying, and backordering cost structure. We consider here the general situation where the supply price associated with a supply market in any time period t may depend, in general, upon the quantity produced at every supply market in every time period. Similarly, the demand price associated with a demand market at any time period may depend upon, in general, the demand for the commodity at every demand market in every time period, that is,

$$\pi = \hat{\pi}(s) \tag{6}$$

and

$$p = \hat{p}(d) \tag{7}$$

where $\hat{\pi}$ and $\hat{p}$ are known smooth functions.

We consider here also the general situation where the cost of transportation, inventorying, and backordering may depend, in general, upon the quantities shipped between every pair of supply and demand markets within every time period, the quantities inventoried at the supply and the demand markets between every pair of successive time periods, and the quantities backordered at every demand market between every pair of time periods, that is,

$$c = \hat{c}(x) \tag{8}$$

where $\hat{c}$ is a known smooth function.

The dynamic spatial price equilibrium conditions have been formulated as a variational inequality in Nagurney and Aronson (1988), following Dafermos (1980), Florian and Los (1982), and Friesz, et.al. (1981). In the framework with gains and losses we can write down immediately the variational inequality formulation.

Theorem: A commodity pattern (s,d,x) satisfying (1), (2) and (3) is in equilibrium if and only if

$$\hat{\pi}(s)\,(s'-s) + \hat{c}(x)\,(x'-x) - \hat{p}(d)\,(d'-d) \geq 0 \tag{9}$$

for all (s',d',x') satisfying (1), (2), and (3).

Hence, the structure of the variational inequality (9) is identical to that in Nagurney and Aronson (1988), but the feasible set defined by constraints (1), (2), and (3) is distinctly different and more complex.

When the supply and demand price and cost functions satisfy the strong monotonicy property

$$[\hat{\pi}(s') - \hat{\pi}(s'')][s'-s''] + [\hat{c}(x') - \hat{c}(x'')][x'-x''] - [\hat{p}(d') - \hat{p}(d'')][d'-d''] \geq \alpha (\| s'-s''\|^2 + \| x'-x''\|^2 + \| d'-d''\|^2) \quad (10)$$

for all (s',d',x'), (s'',d'',x'') satisfying (1), (2), and (3) where α is a positive constant, there exists a unique equilibrium which can be computed by a general iterative scheme devised by Dafermos (1983).

In the special case where the Jacobian matrices $\left[\frac{\partial \hat{\pi}}{\partial s}\right]$, $\left[\frac{\partial \hat{c}}{\partial x}\right]$, and $-\left[\frac{\partial \hat{p}}{\partial d}\right]$ are symmetric it is easy to see that (s,d,x) satisfies (9) if and only if it minimizes the functional

$$F(s,d,x) = \int \hat{\pi}(s)ds + \hat{c}(x)dx - \hat{p}(d)dd \quad (11)$$

for (s,d,x) satisfying (1), (2), and (3). In the symmetric case then there exists a unique equilibrium which can be constructed, at least in principle, by standard convex programming algorithms. For computational comparisons of the Frank-Wolfe (1956) algorithm and computationally efficient equilibration schemes in the framework of the static s.p.e.p. without gains and losses see Nagurney (1987). For computational experience with equilibration schemes by demand markets in time and by supply markets in time in the framework of the dynamic s.p.e.p. without multipliers see Nagurney and Aronson (1988).

We now present three alternative variational inequality formulations of equilibrium conditions (5) equivalent to (9), but defined over Cartesian products of sets, in a manner similar to that given in Nagurney and Aronson (1988) for the dynamic spatial price equilibrium problem without gains and losses. We will then suggest a Gauss-Seidel decomposition procedure for the computation of the equilibrium. Finally, we will give an adaptation of an equilibration scheme for the case of multipliers for the solution of the embedded quadratic programming problems.

We define the vector $\bar{\pi} \in R^{nP}$ with component vectors $\bar{\pi}_{jt'} \in R^{nPjt'} = \{(\pi_{11},...,\pi_{11})\} \in R^{nP11jt'},...\{(\pi_{mT},...,\pi_{mt})\} \in R^{nPmTjt'}$, and the vector $\bar{p} \in R^{nP}$ with the component vectors $\bar{p}_{jt'} \in R^{nPjt'}=\{(a_{1P_{11jt}P_{11}},...a_{nP_{11jt'}P_{11jt'}P_{11}})\} \in R^{nP_{11jt'}},...,$ $\{(a_{1P_{mTjt'}P_{mT}},...a_{nP_{mTjt'}P_{mTjt'}P_{mT}})\} \in R^{nP_{mTjt'}}$, where $iP_{ktjt'}$ denotes the i-th path in the set $P_{ktjt'}$.

Using(1), (2), (3), (4), (6), and (8), we deduce that inequality (9) can be written as

$$\bar{\pi}(y)(y'-y) + C(y)(y'-y) - \hat{p}(d)(d'-d) \geq 0 \qquad \text{for all } (y',d') \in K^1, \quad (12)$$

where $K^1 \equiv \Pi_{j=m+1}^{m+n} \Pi_{t=1}^{T} K_{jt}^1$, where each K_{jt}^1 is given by

$$K^1_{jt} = \{(x_r, r\mathcal{E}P_{jt}, d_{jt}) \mathcal{E} R^{nP_{jt}} \times R \mid x_r \geq 0, d_{jt} = \sum_{r\mathcal{E}P_{jt}} a_r x_r\} . \quad (13)$$

We let y_{jt} denote the vector of path flows for paths contained in P_{jt}.

Similarly, using (1), (2), (3), (4) , and (7), and (8) we deduce that inequality (9) can be written as

$$\hat{\pi}(s)(s'-s) + C(y)(y'-y) - \tilde{p}(y)(y'-y) \geq 0 \qquad \text{for all } (s',y') \mathcal{E} K^2, \quad (14)$$

where $K^2 \equiv \Pi_{i=i}^{m} \; \Pi_{t=1}^{T} \; K^2_{it}$, where K^2_{it} is given by

$$\{(s_{it}, (x_r, r\mathcal{E} P^{it})), \mathcal{E} RxR^{nP_{it}} \mid x_r \geq 0 \text{ and } s_{it} = \sum_{r\mathcal{E}P_{it}} x_r\}. \quad (15)$$

Finally, using (1), (2), (3), (5), (6), and (7) we deduce that inequality (9) can be written as

$$\tilde{\pi}(y)(y'-y) + C(y)(y'-y) - \tilde{p}(y)(y'-y) \geq 0 \qquad \text{for all } y' \mathcal{E} K^3, \quad (16)$$

where $K^3 \equiv \prod_{r\mathcal{E}P} K_r$ and each K_r is given by

$$\{x_r \mid x_r \geq 0\} . \quad (17)$$

3. AN ALGORITHM FOR THE DYNAMIC SPATIAL PRICE EQUILIBRIUM PROBLEM WITH GAINS AND LOSSES

In a recent paper Nagurney and Aronson (1988) outlined Gauss-Seidel type serial linearization decomposition methods for the solution of the dynamic s.p.e.p without gains and losses. They presented decomposition schemes by demand markets in time and by supply markets in time, which were based on similar methods proposed for the static problem in Nagurney (1987). These methods, in combination with equilibration schemes, computed efficiently the equilibrium for problems with more than 10,000 variables. Since the structures of the variational inequalities (12), (14), and (16) governing the equilibrium for the dynamic problem with gains and losses are very closely related to those governing the equilibrium for the problem without (cf. Nagurney and Aronson (1988)), those same methods may be applied with minor appropriate modifications, in view of equations (1) through (4), to solve the more general model.

We first present a decomposition method by demand markets in time and then the equilibration method. Subsequently we provide computational experience on randomly generated large-scale problems with gains and losses.

This algorithm proceeds in a serial manner from time period to time period, solving the demand market subproblem for each demand market in a given time period until variational inequality (12) is solved.

Given a fixed demand market j at time period t, we construct new functions $\tilde{\tilde{\pi}}_{jt}$, new cost functions $\tilde{c}_{it'jt'}$ and $\tilde{C}_{jt}$, and a new demand price function $\tilde{p}_{jt}$, which are linear and are defined as follows:

$$\tilde{\tilde{\pi}}_{jt}(y_{jt}) = D_{\tilde{\pi}_{jt}}(y')y_{jt} + (\tilde{\pi}_{jt}(y') - D_{\tilde{\pi}_{jt}}(y')(y'_{jt})) \quad (18)$$

where $D_{\tilde{\pi}_{jt}}$ denotes the diagonal of the Jacobian of the $\tilde{\pi}_{jt}$ functions with respect to y_{jt}, and y' denotes the latest y, that is, from the previously solved subproblem,

$$\tilde{C}_r = \sum_{it'j't''} a_{(it'j't'')r} \; \tilde{c}_{it'j't''} \; \delta_{(it'j't'')r}, \text{ for } r \in P_{jt} \tag{19}$$

where

$$\tilde{c}_{it'j't''} = D_{c_{it'j't''}}(x') \, x_{it'j't''} + (c_{it'j't''}(x'_{it'j't''}) - D_{c_{it'j't''}}(x) \, x'_{it'j't''}) \tag{20}$$

where $D_{c_{it'j't''}}$ denotes the diagonal of the Jacobian of the cost function $c_{it'j't''}$ with respect to $x_{it'j't''}$, where $x_{it'j't''}$ is given by (3), and

$$\tilde{p}_{jt}(d_{jt}) = D_{p_{jt}}(d')d_{jt} + (p_{jt}(d') - D_{p_{jt}}(d') \, d'_{jt}) \tag{21}$$

where $D_{p_{jt}}$ denotes the t-th diagonal element of the Jacobian of the demand price function p_{jt} and d' denotes the vector of the latest available demands.

One then solves the decomposed variational inequality subproblem

$$\sum_{it',r \in P_{it'jt}} \tilde{\pi}_{it'jt}(x_r)(x_r - x_r) + \sum_{r \in P_{jt}} \tilde{C}_r(y_{jt})(x'_r - x_r) - \tilde{p}_{jt}(d_{jt})(d'_{jt} - d_{jt}) \geq 0$$

$$\text{for all } (y_{jt}, d_{jt}) \in K^1_{jt} \tag{22}$$

where $\tilde{\pi}_{it'jt}$ denotes the it'-element of the vector $\tilde{\pi}_{jt}$.

But due to the construction of $\tilde{\pi}_{jt}$, $\tilde{C}_{jt}$, $\tilde{p}_{jt}$, (22) is equivalent to the solution of the quadratic programming problem

$$\text{Min} \sum_{it',r \in P_{it'jt}} \int_0^{x_r} \tilde{\pi}_{it'jt}(x)dx + \sum_{it'j't'} \int_0^{\sum_{r \in P_{jt}} x_{r(it'j't'')}} \tilde{c}_{it'j't''}(z)dz - \int_0^{d_{jt}} \tilde{p}_{jt}(y)dy \tag{23}$$

over K^1_{jt} which can be effectively solved by a demand market in time equilibration scheme for multipliers we present in Section 4. Note that this quadratic programming problem differs from the one that is solved when the multipliers in (1) are equal to one.

We initiate the algorithm with a $(y',d') \in K^1$. We then form the functions $\tilde{\pi}_{11}$, $\tilde{C}_r, r \in P_{11}$, and $\tilde{p}_{11}$ and solve (23) for j=t=1. We continue in this manner through demand market m in period T. This is termed an iteration.

In a similar manner, one may construct a decomposition method by supply markets in time for the solution of variational inequality (14).

4. A DEMAND MARKET EQUILIBRATION OPERATOR FOR GAINS AND LOSSES

In this section we describe an algorithm for the solution of the dynamic spatial price equilibrium problem with gains and losses in which the supply price, demand price, and cost functions are linear. The supply price at a supply market in a period t depends only upon the quantity of the commodity produced at the supply market in the period t. The demand price at a demand market in a period t depends only upon the quantity of the commodity demanded at the demand market in the period t. The inventorying cost between two periods t and t+1 at a supply (demand) market depends only upon the quantity of the commodity inventoried at the supply (demand) market between the two successive time periods. The backordering cost at a demand market between two time periods depends only upon the quantity of the commodity backordered between the two time periods. The transportation cost, in turn, between a pair of supply and demand markets within a time period depends only upon the quantity of the commodity shipped between that pair of supply and demand markets within that time period.

Hence, we assume that the supply price functions (6) and the demand price functions (7) are of the form

$$\pi_{it} = \hat{\pi}_{it}(s_{it}) = r_{it}s_{it} + u_{it}\ ;\ \ i=1,...,m;\ \ t=1,...,T \tag{24}$$

and

$$p_{jt} = \hat{p}_{jt}(d_{jt}) = -m_{jt}d_{jt} + q_{jt}\ ;\ \ j=1,...,n;\ \ t=1,...,T \tag{25}$$

where $r_{it}, m_{jt}, u_{it}, q_{it} > 0$ for all i, j and t. The cost functions (8) are also linear and of the form

$$c_{itjt'} = \hat{c}_{itjt'}(x_{itjt'}) = g_{itjt'}\,x_{itjt'} + h_{itjt'}\ ;\quad \begin{matrix} i=1,...,m;\ \ t=1,...,T, \\ j=1,...,n;\ \ t'=1,...,T \end{matrix} \tag{26}$$

with $g_{itjt'}$ and $h_{itjt'} > 0$ where $x_{itjt'}$ satisfies (1) and (2).

Following Samuelson (1952), Takayama and Judge (1971), and Florian and Los (1982), this model has an equivalent optimization formulation with objective function

$$\text{Min} \sum_{it} \int_0^{s_{it}} \hat{\pi}_{it}(x)dx + \sum_{itjt'} \int_0^{\sum_r x_{r(itjt')}} \hat{c}_{itjt'}(z)dz - \sum_{jt'} \int_0^{d_{jt'}} \hat{p}_{jt'}(y)dy \tag{27}$$

subject to constraints (1), (2), and (3) or equivalently,

$$\text{Min} \sum_{it} \int_0^{\sum_{r\in P_{it}} x_r} \hat{\pi}_{it}(x)dx + \sum_{itjt'} \int_0^{x_{itjt'}} \hat{c}_{itjt'}(z)dz - \sum_{jt'} \int_0^{\sum_{r\in P_{jt'}} a_r x_r} \hat{p}_{jt'}(y)dy. \tag{28}$$

We now present an adaptation of an equilibration operator proposed in Nagurney and Aronson (1988) for the solution of the dynamics s.p.e.p. without multipliers and conceived by Dafermos and Sparrow (1969) for the traffic network equilibrium problem with fixed demands. As in that dynamic s.p.e.p., in the case of linear supply price, demand price, and cost functions, this operator induces an algorithm which takes on a simple and elegant form for computational purposes. The operator is associated with the

demand markets. Starting from an initial commodity flow pattern satisfying (1), (2), and (3) we construct a sequence of feasible flow patterns which reduce the value of the objective function OF given by (28). For convergence results see Dafermos and Sparrow (1969) and Nagurney (1987).

This algorithm proceeds from time period to time period and from demand market to demand market within a time period, equilibrating the sum of the supply price and the effective cost of the path with the value of the commodity at the demand market if there is positive flow from the supply market to the demand market until equilibrium conditions (5) hold for each demand market.

We seek an operator M_{jt} to demand market j in period t.

Let $y \in R^{nP} = \{x_r \mid x_r \geq 0\}$. We define $y' = M_{jt}$ as follows:

We define r_{max} and r_{min} and $(it')_{max}$ and $(it')_{min}$ by

$$\pi_{(it')_{min}} + C_{r_{min}} - p_{jt} = \min\{\pi_{it'} + C_r - a_r p_{jt} : \quad i = 1,...,m;\ t'=1,...,T,\ r \in P_{it'jt}\} \tag{29}$$

$$\pi_{(it')_{max}} + C_{r_{max}} - p_{jt} = \max\{\pi_{it'} + C_r - a_r p_{jt} : \quad i = 1,...,m;\ t'=1,...,T,\ r \in P_{it'jt},\ x_r > 0\}\ . \tag{30}$$

The calculation then of $y'=M_{jt}y$ consists of the calculation of two new commodity flows $x'_{r_{max}}$ and $x'_{r_{min}}$ while holding all other commodity flows fixed. Let $\Delta x_{r_{max}}$ denote $x'_{r_{max}} - x_{r_{max}}$ and $\Delta x_{r_{min}} = x'_{r_{min}} - x_{r_{min}}$. Equating then $(\pi_{(it')_k} + C_{r_k} - a_{r_k} p_{jt})$ for k=max and k=min at the new commodity flow y' to zero we obtain the 2x2 system of equations

$$\begin{aligned}&\Big(\sum_{itjt'} g_{itjt'}\, \delta_{(itjt')r_{max}}\, a^2_{(itjt')r_{max}}\, m_{jt}\ a^2_{r_{max}} + r_{(it)_{max}}\Big)\Delta x_{r_{max}} \\ &+ \Big(\sum_{itjt'} g_{itjt'}\, \delta_{(itjt')r_{max}}\, \delta_{(itjt)r_{min}}\, a_{(itjt')r_{min}}\, a_{(itjt')r_{max}} \\ &+ m_{jt} a_{r_{min}}\, a_{r_{max}} + r_{(it)_{max}}\, \delta_{(it)_{max}}\, \delta_{(it)_{min}}\Big)\Delta x_{r_{min}} = b_1\end{aligned} \tag{31}$$

$$\begin{aligned}&\Big(\sum_{itjt'} g_{itjt'}\, \delta_{(itjt')r_{max}}\, \delta_{(itjt')r_{min}}\, a_{(itjt')r_{max}}\, a_{(itjt')r_{min}} + m_{jt} a_{r_{max}} a_{r_{min}} \\ &+ r_{(it)_{min}}\, \delta_{(it)_{max}}\, \delta_{(it)_{min}}\Big)\Delta x_{r_{max}} + \Big(\sum_{itjt'} g_{itjt'}\, \delta_{(itjt')r_{min}}\, a^2_{(itjt')r_{min}} \\ &+ m_{jt} a^2_{r_{min}} + r_{(it)_{min}}\Big)\Delta x_{r_{min}} = b_2\end{aligned} \tag{32}$$

where $\delta_{(it)_{max}}\, \delta_{(it)_{min}} = 1$, if $(it)_{max} = (it)_{min}$, and 0 otherwise, and

$$\begin{aligned}b_1 = &\Big(-\sum_{itjt'} g_{itjt'}\, x_{itjt'}\, \delta_{(itjt')r_{max}} - h_{itjt'}\, \delta_{(itjt')r_{max}}\Big)\, a_{(itjt')r_{max}} \\ &+ (-m_{jt} d_{jt} + q_{jt}) a_{r_{max}} - r_{(it)_{max}}\, s_{(it)_{max}} - u_{(it)_{max}}\ ,\end{aligned} \tag{33}$$

$$+ (-m_{jt}d_{jt} + q_{jt})a_{r_{max}} {}^{-r}(it)_{max} {}^{s}(it)_{max} - {}^{u}(it)_{max}, \tag{33}$$

$$b_2 = (-\sum_{itjt'} g_{itjt'} x_{itjt'} \delta_{(itjt')r_{min}} - h_{itjt'} \delta_{(itjt')r_{min}}) a_{(itjt')r_{min}}$$
$$+ (-m_{jt}d_{jt} + q_{jt})a_{r_{min}} {}^{-r}(it)_{min} {}^{s}(it)_{min} - {}^{u}(it)_{min} \tag{34}$$

where equations (1), (2), and (3) have been used for simplicity.

Under our assumptions,

$$A \equiv \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix}$$

where α_{11} and α_{12} denote the coefficients of $\Delta x_{r_{max}}$ and $\Delta x_{r_{min}}$ in (31), and α_{21} and α_{22} denote the coefficients of $\Delta x_{r_{max}}$ and $\Delta x_{r_{min}}$ in (32), respectively. A is nonsingular because $\lambda = \det(A)$ is greater than zero. Hence, an application of Cramer's rule to (31) and (32) yields:

$$\Delta x_{r_{max}} = \frac{\det \begin{matrix} b_1 & \alpha_{12} \\ b_2 & \alpha_{22} \end{matrix}}{\lambda} \qquad \Delta x_{r_{min}} = \frac{\det \begin{matrix} \alpha_{11} & b_1 \\ \alpha_{21} & b_2 \end{matrix}}{\lambda}$$

Since both $x'_{r_{max}}$ and $x'_{r_{min}}$ must be ≥ 0, we must assure feasibility and, at the same time, improve the value of OF. If the computed values of $\Delta x_{r_{max}} \in [-x_{r_{max}}, \infty)$ and $\Delta x_{r_{min}} \in [-x_{r_{min}}, \infty)$, feasibility of the improved solution is assured. If $\Delta x_{r_{max}} \in (-\infty, -x_{r_{max}})$ and $\Delta x_{r_{min}}$ $(-\infty, -x_r)$, we set $\Delta x_{r_{max}} = -x_{r_{max}}$ and $\Delta x_{r_{min}} = -x_{r_{min}}$. On the other hand , if the computed $\Delta x_{r_{max}} \in [-x_{r_{max}}, \infty)$, but $\Delta x_{r_{min}} \in (-\infty, -x_{r_{min}})$, then set $\Delta x_{r_{min}} = -x_{r_{min}}$ and retain the computed $\Delta x_{r_{max}}$. It is easy to verify that since $b_1 < b_2$, in this case $\Delta x_{r_{max}} \in [-x_{r_{max}}, 0)$. Finally, if $\Delta x_{r_{max}} \in (-\infty, -x_{r_{max}})$, but $\Delta x_{r_{min}} \in [-x_{r_{min}}, \infty)$, we proceed as follows: If $\Delta x_{r_{min}} \in [-x_{r_{min}}, 0]$, we set $\Delta x_{r_{max}} = -x_{r_{max}}$ and retain $\Delta x_{r_{min}}$. If $\Delta x_{r_{min}} \in (0, \infty)$, we compute the solution to equation (32) for $\Delta x_{r_{min}}$, with $\Delta x_{r_{max}} = -x_{r_{max}}$. If the value of $-b_1$ at this proposed x' is ≥ 0, we retain these values. Otherwise, set $\Delta x_{r_{min}} = 0$, and solve for $\Delta x_{r_{max}}$ using (31). If $\Delta x_{r_{max}} < -x_{r_{max}}$, set $\Delta x_{r_{max}} = -x_{r_{max}}$.

Note that by construction, $M_{jt}y=y$ if and only if for r_{max} and r_{min} and jt equilibrium conditions (5) hold. In particular, $\pi_{(it')max} + C_{r_{max}} - a_{r_{max}} p_{jt} = \pi_{(it')min} + C_{r_{min}} - a_{r_{min}} p_{jt} = 0$, where $x_{r_{max}} > 0$ and r_{max} and r_{min} have been chosen so that for any i and t and $r \in P_{jt}$

$$\pi_{(it')min} + C_{r_{min}} - a_{r_{min}} p_{jt} \leq \pi_{it} + C_r - a_r p_{jt} \leq \pi_{(it')max} + C_{r_{max}} - a_{r_{max}} p_{jt}. \quad (37)$$

Therefore, if $M_{jt}y=y$, then all supply markets at all time periods, with $x_r > 0$ satisfy (5).

Also, it is clear that M_{jt} is a continuous mapping from R^{nP} to R^{nP}. Finally, M_{jt} decreases OF and $OF(M_{jt}y) = OF(y)$ for some $y \in R^{nP}$ implies that $M_{jt}y=y$.

We thus define the operator $M^{(1)}$ as the composition of the operators $M_{11}... M_{nT}$. Hence $M^{(1)}$ has the same properties as M_{jt}. In the case of all multipliers being identically one this operator collapses to the demand market equilibration operator given in Nagurney and Aronson (1988).

5. COMPUTATIONAL EXPERIENCE WITH AN EQUILIBRATION OPERATOR FOR GAINS AND LOSSES

Here we consider the dynamic spatial price equilibrium problem where the supply price functions are given by (24), the demand price functions are given by (25), and the inventorying, backordering, and transportation cost functions are given by (26) and we give computational results for the algorithms outlined in Section 4 for the solution of such dynamic s.p.e.p.'s.

All of the examples in this section were generated as follows. (See Nagurney and Aronson (1988)) The supply price, demand price, and transportation cost function slopes and intercepts (cf, (24), (25), (26)), were generated randomly and uniformly as whole numbers within the following ranges: $r_{it} \in [3,10]$, $u_{it} \in [10,15]$, $-m_{jt} \in [-1,-5]$, $q_{jt} \in [150,650]$, $g_{itjt} \in [1,15]$, and $h_{itjt} \in [10,25]$, i=1,...,m; j=m+1,...,m+n; t=1,...,T. The supply price inventorying cost slopes g_{itit+1} and the intercepts h_{itit+1} were generated within the ranges defined by .075 times the lower and upper limits of the supply price cost function ranges, respectively. The demand inventorying cost slopes g_{jtjt+1} and the backordering slopes g_{itjt-1}, were generated in the range defined by .075 times the sum of the lower limits for the supply price and transportation cost function slopes as the lower limit and .075 times the sum of the upper limits of the slopes for the supply price and transportation cost functions as the upper limit. The intercepts h_{jtjt+1} and h_{jtjt-1} were generated in a similar fashion, utilizing the sum of the supply price and transportation cost intercept limits. The multipliers $a_{itjt'}$ were generated in the range [.95,.99], the backordering multipliers were set equal to 1.

The initial commodity flow $y^{(1)}$ was generated as follows. The flow on paths corresponding to transportation links were generated whole numbers in the range 1 through 5, and identical for a given example; all other flows were set equal to zero.

The equilibration operator $M^{(1)}$ proposed here was coded in FORTRAN and all examples were run on the IBM 3084QX under the VM operating system at the Cornell University Production Supercomputer Facility, except where noted. The termination criterion was:

$| \pi_{it} + C_r - a_r p_{jt'} | \leq 10$ for all supply and demand market pairs and time periods, and for all paths $r \in P_{itjt'}$, such that $x_r > 0$, and $| \min(\pi_{it} + C_r - a_r p_{jt'}) | \leq 10$ otherwise. The number of iterations and CPU time were measured (exclusive of data generation, input, setup and output times) and reported for all the examples.

We first considered examples in which only inventorying at supply markets is allowed. In this case (cf. Figure 1), the number of links in the network representation is $mnT + m(T-1)$. For these examples, reported in Table 1, we fixed the number of supply and demand markets and increased the number of time periods incrementally by two periods, starting with two time periods and continuing through ten time periods. The examples are identical, except for the multipliers, to the examples of Table 1 in Nagurney and Aronson (1988).

For the purpose of comparison we also reported in Table 1 the CPU times and the number of iterations for the examples on the CDC CYBER 830 at the University of Massachusetts, Amherst which was the system used for the computation of the equilibrium in Nagurney and Aronson (1988). As can be seen from Table 1 the new system is faster by an order of magnitude. The examples with multipliers required no more iterations (in most cases several fewer) than the number required to solve the problem without gains and losses (see Nagurney and Aronson (1988) Table 1). This may be due, in part, to the initialization procedure.

The data demonstrate that given the state-of-the-art in computer technology, problems with as many as 10,000 variables can now be solved in CPU seconds, rather than minutes. We note that the number of iterations required to solve a given problem on the IBM machine would differ, in some cases, from the number required on the CDC machine; this is due to different word sizes on the two machines.

We then considered a series of examples in which inventorying at both supply and demand markets is permitted and backordering is also allowed and we varied the time periods from two through five. This is the network model given in Figure 1. Our results are reported in Table 2. These examples, with the exception of the multipliers, are identical to those in Table 3 in Nagurney and Aronson (1988).

Table 1: Computational Experience for the Equilibration Method $M^{(1)}$ for Randomly Generated Dynamic S.P.E. Problems with Gains and Losses
CPU time in seconds (# of iterations)*

m	n		T= 2	4	6
10	10	$M^{(1)}$	0.91(8)[1.1(8)]	.416(9)[5.1(11)]	1.02(10)[12.5(11)]
10	20	$M^{(1)}$	.266(11)[3.1(10)]	1.31(14)[15.6(14)]	3.26(15)[35.9(14)]
20	20	$M^{(1)}$	.761(9)[8.9(11)]	4.35(13)[46.6(14)]	11.65(13)[114.7(14)]

m	n		T= 8	10
10	10	$M^{(1)}$	2.11(10)[25.4(11)]	3.72(11)[37.4(10)]
10	20	$M^{(1)}$	6.31(14)[71.4(14)]	10.86(15)[109.5(14)]
20	20	$M^{(1)}$	23.16(14)[220.9(15)]	90.5(16)[351.4(14)]

*Note the CPU time and the number of iterations required on the CDC CYBER 830 is given in [ ].

Table 2: Computational Experience for the Equilibration Method $M^{(1)}$ for Randomly Generated Dynamic S.P.E. Problems with Gains and Losses Inventorying at Supply Markets, at Demand Markets, and Backordering
CPU time in seconds (# of iterations)

m	n	T=	2	3	4	5
10	10	$M^{(1)}$	.3(12)	1.1(15)	2.7(15)	6.1(14)
10	15	$M^{(1)}$	.5(13)	1.5(13)	4.0(14)	9.5(15)
15	15	$M^{(1)}$	.9(11)	3.6(16)	8.1(14)	18.3(14)
15	20	$M^{(1)}$	1.5(14)	5.0(15)	14.1(19)	31.8(20)
20	20	$M^{(1)}$	2.5(13)	8.8(14)	23.3(19)	50.4(19)

6. COMPUTATIONAL EXPERIENCE WITH A DECOMPOSITION SCHEME FOR GAINS AND LOSSES

In this section we consider the general dynamic spatial price equilibrium model outlined in Section 2. Our computational experience is for examples with linear asymmetric functions. Hence, we assume here that the supply price functions (6) are given by

$$\pi_{it} = \hat{\pi}_{it}(s) = \sum_{jt'} r_{itjt'} s_{jt'} + u_{it}, \tag{38}$$

the demand price functions (7) are given by

$$P_{jt} = \hat{p}_{jt}(d) = -\sum_{it'} m_{jtit'} d_{it'} + q_{jt}, \tag{39}$$

and the cost functions (8) are given by

$$c_{itjt'} = \hat{c}_{itjt'}(x) = \sum_{jt'} g_{itjt'} x_{itjt'} + h_{itjt'} \tag{40}$$

where the not necessarily symmetric Jacobians of the supply price, demand price, and cost functions are positive definite. We refer, henceforth, to the Gauss-Seidel decomposition by demand markets in time $GS^{(1)}$. We refer to the $GS^{(1)}$ method with the embedded $M^{(1)}$ method as $GS^{(1)}M^{(1)}$.

All the examples were generated as follows. The number of cross-terms for any supply price, demand price or cost function (cf. (38), (39), (40)) ranged from 1 to 5 and were generated to ensure that the Jacobian matrices of these functions were strictly diagonally dominant and, hence, positive definite. The diagonal terms, the intercepts and the initial commodity pattern were generated in the manner outlined in Section 5 as were the multipliers.

These algorithms were also coded in FORTRAN and all examples run on the IBM 3084QX at the Cornell University Production Supercomputer Facility. The termination criterion used was that the condition for termination of the equilibration operators given in Section 5 had to hold for two consecutive iterations of the Gauss-Seidel scheme.

Parallel to the computational tests of Section 5, we again considered first examples in which inventorying only at the supply markets is allowed. These examples, with the exception of the multipliers, are identical to those in Table 4 in Nagurney and Aronson (1988). The results are reported in Table 3.

Finally, we considered a series of examples, in which inventorying is permitted at both supply and demand markets and backordering is also allowed and varied the time periods from two through five. These experiments are given in Table 4. These examples are identical to those in Table 6 in Nagurney and Aronsson (1988), except here we also have the multipliers.

Table 3: Computational Experience for the Decomposition Method $GS^{(1)}M^{(1)}$ for Randomly Generated Dynamic S.P.E. Problems with Gains and Losses Inventorying at Supply Markets
CPU time in seconds (# of iterations)

m	n		T= 2	4	6	8	10
10	10	$GS^{(1)}M^{(1)}$	.1(7)	.7(9)	1.6(9)	3.1(9)	6.0(10)
10	20	$GS^{(1)}M^{(1)}$	.3(9)	1.9(13)	4.3(12)	8.2(12)	14.9(12)
20	20	$GS^{(1)}M^{(1)}$	.8(10)	4.8(12)	12.0(12)	25.9(15)	52.2(15)

Table 4: Computational Experience for the Decomposition Method $GS^{(1)}M^{(1)}$ for Randomly Generated Dynamic S.P.E. Problems with Gains and Losses Inventorying at Supply Markets, at Demand Markets, and Backordering
CPU time in seconds (# of iterations)

m	n		T= 2	3	4	5
10	10	$GS^{(1)}M^{(1)}$	.4.(10)	2.3(14)	4.8(12)	9.1(10)
10	15	$GS^{(1)}M^{(1)}$	.7(11)	3.9(18)	8.6(14)	17.2(14)
15	15	$GS^{(1)}M^{(1)}$	1.2(10)	4.4(11)	14.3(15)	27.8(13)
15	20	$GS^{(1)}M^{(1)}$	1.7(10)	6.3(11)	22.6(17)	43.6(14)
20	20	$GS^{(1)}M^{(1)}$	2.8(12)	12.0(14)	35.0(18)	56.3(13)

7. CONCLUSIONS

We have presented a general dynamic finite horizon spatial price equilibrium model with gains and losses which supersedes both earlier dynamic and static models. We gave the governing equilibrium conditions and provided alternative variational inequality formulations. We applied a Gauss-Seidel type decomposition method by demand markets in time, in which we embedded an equilibration multiplier method, for the solution of a variety of problem scenarios. Our computational experience demonstrates that given the state-of-the-art of computer technology, problems on the order of thousands of variables can be solved in only several seconds of CPU time.

ACKNOWLEDGEMENTS

This research was conducted on the Cornell University Production Supercomputer Facility of the Center for Theory and Simulation in Science and Engineering, which is funded, in part, by the National Science Foundation, New York State and IBM Corporation.

It was also supported by a Faculty Research Grant from the University of Massachusetts and a summer grant from the School of Management. The author would like to thank Roberto Sanz de Santa Maria for assisting with the numerical calculations presented in this paper and Jay Aronson for many stimulating conversations.

REFERENCES

Dafermos S., 1980, "Traffic Equilibrium and Variational Inequalities", *Transportation Science*, 14:42-54.

Dafermos, S., 1983, "An Iterative Scheme for Variational Inequalities", *Mathematical Programming*; 26:40-47.

Dafermos, S. and A. Nagurney, 1984, "Sensitivity Analysis for the General Spatial Economics Equilibrium Problem", *Operations Research*; 32:1069-1086.

Dafermos, S. and A. Nagurney, 1987, "Oligopolistic and Competitive Behavior of Spatially Separated Markets", *Regional Science and Urban Economics*, 17:245-254.

Dafermos, S. and F.T. Sparrow, 1969, "The Traffic Assignment Problem for a General Network", *Journal of Research of the National Bureau of Standards*, 73B, no. 2:91-118.

Florian, M. and M. Los, 1982, "A New Look at Static Spatial Price Equilibrium Models", *Regional Science and Urban Economics*, 12:579-597.

Frank, M. and P. Wolfe, 1956, "An Algorithm for Quadratic Programming", Naval *Research Logistics Quarterly*, 3:95-110.

Friesz, T., P.T. Harker, and R.L. Tobin, 1984, "Alternative Algorithms for the General Network Spatial Price Equilibrium Problem", *Journal of Regional Science*, 24:473-507.

Friesz, T.L., R.L. Tobin, and P.T. Harker, 1981, "Variational Inequalities and Convergence of Diagonalization Methods for Derived Demand Network Equilibrium Problems", Report CUE-FNEM-1981-10-1, Department of Civil and Urban Engineering, University of Pennsylvania.

Friesz, T.L., R.L. Tobin, T.E. Smith, and P.T. Harker, 1983, "A Nonlinear Complementarity Formulation and Solution Procedure for the General Derived Demand Network Equilibrium Problem", *Journal of Regional Science*, 23:337-359.

Guise, J.W.B., 1979, "An Expository Critique of the Takayama-Judge Models of Interregional and Intertemporal Market Equilibrium", *Regional Science and Urban Economics*, 9:83-95.

Jones, P.C., R. Saigal, and M. Schneider, 1984a, "Computing Nonlinear Network Equilibria", *Mathematical Programming*, 31:57-66.

Jones, P.C., R. Saigal, and M. Schneider, (1984b), "A Variable Dimension Homotopy for Computing Spatial Equilibria", *Operations Research Letters* 3:19-24.

Judge, G.G. and T. Takayama, (eds.), 1973, *Studies in Economic Planning Over Space and Time*, North-Holland, Amsterdam.

Nagurney, A., 1984, "Comparative Tests of Multimodal Traffic Equilibrium Methods", *Transportation Research*, 18B:469-485.

Nagurney, A., 1987, "Computational Comparisons of Spatial Price Equilibrium Methods", *Journal of Regional Science*, 27:55-76.

Nagurney, A. and J. Aronson, 1988, "A General Dynamic Spatial Price Equilibrium Model: Formulation, Solution, and Computational Results", *Journal of Computational and Applied Mathematics* 22, in press.

Pang, J-S., 1984, "Solution of the General Multicommodity Spatial Equilibrium Problem by Variational and Complementarity Methods", *Journal of Regional Science*,. 24:403-414.

Pang, J-S., 1985, "Asymmetric Variational Inequality Problems Over Product Sets: Applications and Iterative Methods", *Mathematical Programming*, 31:206-219.

Samuelson, P.A., 1952, "A Spatial Price Equilibrium and Linear Programming", *American Economic Review*, 42:283-303.

Samuelson, P.A., 1957, "Intertemporal Price Equilibrium: A Prologue to the Theory of Speculation", *Weltwirtschaftliches Archiv*, 79:181-219.

Takayama, T., H. Hashimoto, and N.P. Uri, 1984, "Spatial and Temporal Price and Allocation Modeling: Some Extensions", *Socio-Economic Planning Science*, 18:227-234.

Takayama, T. and G.G. Judge, 1964, "Equilibrium Among Spatially Separated Markets: A Reformulation", *Econometrica*, 32:510-524.

Takayama, T. and G.G. Judge, 1971, *Spatial and Temporal Price and Allocation Models*, North-Holland, Amsterdam.

Takayama, T. and N.P. Uri, 1983, "A Note on Spatial and Temporal Price and Allocation Modeling", *Regional Science and Urban Economics*, 13:455-470.

Thore, S., 1986, "Generalized Network Spatial Equilibrium: The Deterministic and Chance-Constrained Case", *Papers of the Thirty-second North American Meetings of the Regional Science Association*, 59:93-102.

Advances in Spatial Theory and Dynamics
A.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 16

Infrastructure and Economic Transformation

T.R. Lakshmanan

1. INTRODUCTION

Since World War II the extensive literature on economic growth and development has emphasized the key role of 'infrastructure' or 'economic and social overhead capital' in national and regional development (Rosenstein-Rodan, 1943; Nurske, 1953; Hirschman, 1963; Nadiri, 1970). This literature attributes a two-fold return to the nation or the region from the provision of social overhead capital. First, the delivery of improved education, health care, and recreation directly augments the welfare of human resources or individuals in the form of better skills, reduced absenteeism, etc. Second, the provision of roads, airports, utilities, etc., improves the productivity of producer capital (machinery, equipment, livestock, etc.) and of consumer capital (housing and residential structures). In the long run the impact of infrastructure creation will be such as to lead to extensive modification in the relative prices both of factors of production and final products. The expectation is that there will emerge a new general equilibrium of costs and prices at a higher level of income and employment.

There has been much discussion of this nature of the crucial role of infrastructure and the temporal order in which infrastructure and other forms of capital should be provided so as to stimulate economic development. However, what infrastructure consists of is rarely explored, and its attributes and composition are largely left to be defined in an ad hoc fashion. Further, while there is a variety of ideas and suggestions about the origins and the processes of economic change and development, there is no coherent theory of economic development into which infrastructure can be incorporated. This makes the analysis of infrastructure more difficult, particularly since countries in different stages of technological evolution are interested in the scope and limitations of infrastructure investment policy in furthering economic development.

In developing countries and in lagging regions of industrialized societies, the level and quality of services flowing from existing infrastructure are often low; in many cases they fall below the 'threshold' levels where their effects on economic output are negligible (a type of low level equilibrium trap). The thrust of development policy in such areas is to promote the transition from a predominantly agricultural and natural resource-based economy to a goods-producing economy. It is in the management of this technical change towards industrialization that infrastructure investments are viewed as playing an important role. Since infrastructure investments are usually large and lumpy, it is important to understand the nature of the contribution that such investments make to national or regional income in countries with scarce resources. The critical analytical questions regarding the promotion of economic development are: Where and when should infrastructure investments be made, and in what form and quantity?

In the affluent industrialized economies, on the other hand, there is a *different* structural transformation underway: a transformation from a goods producing economy towards a dominant service economy - a transformation relating to both *what* is produced, *how* it is produced and *where* (Gershuny, 1978; Stanbeck et.al., 1981). Not only is there a trend towards a greater variety of services but increasingly services are produced jointly with goods. There is a significant growth in producer services and producer service-like functions and an expanding emphasis on investments in human capital. The transformation in the way in which production is organized reflects at once the shift in technology, in labor and consumer markets, and in the organizational basis - the process of service delivery itself becoming increasingly routinized, standardized and 'industrialized' (Levitt, 1976; Gershuny, 1978). As a consequence of the above changes, different types of services are locating at different levels of the urban hierarchy and thus transforming the urban system (Daniels, 1985).

The key attribute of these growing service sector industries is that they are information intensive. They depend on various telecommunications systems not only to solicit business but also to deliver their products. Communication systems are to service industries what roads, railways and canals are to (goods producing) manufacturing. A major effect of the emerging innovations in telecommunications, electronics and computing is to increase the sizes of the service markets by breaking down the market barriers, integrating dispersed markets and facilitating the creation of new markets. Further, these innovations increase the speed, density, and quality of information flows, which in turn augment the potential pace of technological change and the diffusion of innovations. Further, since these developments in the telecommunications sector are taking place in a period of increasing internationalization of the service sector, the facilities and networks extend beyond national boundaries (Undersea Cables, Geostationary Satellites, etc.). Currently the impacts of these telecommunications developments are keenly felt in many information-rich producer services, whose range and quality are being transformed. When the potential of these developments is realized by consumer services as well, major impacts on the range and quality of services, on labor utilization, and work organization are likely.

The key analytical questions here are: What role does this increase in capacity and lowered unit costs in telecommunications and information technology have on future economic growth and in facilitating the transition to a dominant service economy? Given the rapid technological innovation and the growing deregulation of the telecommunications industry, public policy choices on types, sizes and locations of communications infrastructure investments become important to future economic growth in an increasingly international production system.

The objective of this paper is to make a case for a broader and deeper approach to the study of the role of infrastructure in facilitating the economic transformations taking place both in newly industrializing regions and in industrialized societies. The paper begins with a clarification of the term 'infrastructure', a term somewhat loosely used in the literature concerning economic growth and development. The paper then proceeds to a review of the available frameworks for an analysis of the role of infrastructure in facilitating the economic changes noted above. The scope and adequacy of the various recent analyses of the contributions of infrastructure to industrial growth in developing nations and lagging regions in affluent societies are assessed in this regard.

We then proceed to a brief delineation of the scope and dimensions of the change in affluent societies towards a dominant service economy and the concomittant growth of information activities. This rapid evolution of information activities and their incorporation as crucial inputs to production has been facilitated by the spectacular growth of the communications infrastructure.

In Section 5 we attempt to relate the growth of information infrastructure to the economic transformation underway. For this purpose, we review various existing analytical frameworks for specifying the production and consumption of goods and

services and note the inadequacy of models of service production - which involve processing of materials as well as information - many of which display "Coproduction" (producers and consumers jointly producing services).

The need for an interdisciplinary systems approach in order to relate information and communication infrastructures to the broader socioeconomic system is deemed necessary. We present the outline of such a framework, which is amenable to further development and analysis in both static and dynamic modes.

Our objective in calling for a broad perspective and a theoretical reconceptualization of the ongoing economic transformation and the role of infrastructure in such a process is not to take the stance of a theorist and soar off into a visionary state of abstraction, accompanied by formalisms presented in the Greek alphabet. Since existing analytical structures are unequal to the task of representing the ongoing organizational and spatial structural change in the service sector, even an avowed novice in theorizing cannot resist the temptation to identify some basic elements of such a reconceptualization of the industrialized economies.

2. WHAT IS INFRASTRUCTURE?

The concept of infrastructure or overhead capital has been in use since the second World War - often employed in a loose impressionistic manner. Youngson (1967) surveys the evolution of this concept in the hands of early development theorists such as Rosenstein-Rodan (1943), Nurske (1952), and Hirschman (1958). He notes that in Nurske's original statements about infrastructure, (which many other authors have since used as a basis for their own work), the list of overhead capital items is neither consistent nor compatible with his criteria about them. For instance, the characteristics of infrastructure Nurske specified ("Cannot be imported from abroad", "large and costly installations", "provide services basic to any production capacity") are more casual than precise. Other scholars have worked with the same criteria while adding more items to the list of infrastructure investments.

Youngson (1967) rescues the concept of infrastructure from the somewhat woolly thinking of the early scholars through a clarification of the relationships between the economic theory of external (technological and pecuniary) economies and the desired attributes of infrastructure. He concludes that infrastructure is *not a set of things but a set of attributes*. Some or all or none of these attributes may reside in the capital instruments. To the degree the latter possess the attributes, they could be regarded as infrastructure. Two such attributes - both of which accord with the notions of Rosenstein-Rodan and Nurske - can be recognized. First, capital can be viewed as infrastructure to the extent (a) it is a source of external economies and (b) it has to be provided in large units, "ahead of demand". Capital expenditures satisfying either of these attributes should be viewed as infrastructure. Both imply the desirability of a certain amount of public investment (since the pattern of investment in a private enterprise economy, given the external effects, tends not to be socially ideal). The second criterion of provision ahead of demand is truly an expost argument (satisfactory when the outcome is known). On the other hand, there is uncertainty and imperfect knowledge in an economy undergoing transformation and faith in the future is an important factor, according to Nurske (Youngson, 1967).

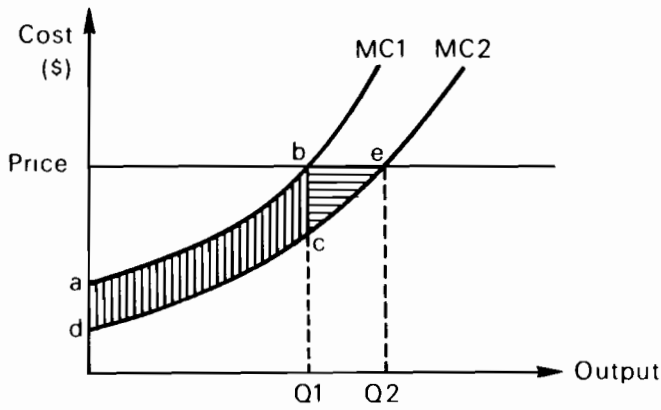
The argument for such infrastructure is particularly strong in the case of those investments which may be thought of as somewhat *nonspecific* in character - that is, those which can be utilized in the production of a wide variety of final outputs such as overhead capital investment in education. The ultimate return to society from education may be out of all proportion to the costs. The indirect benefits which are derived from public and private spending on education extend far beyond the direct benefits (e.g., the economic returns from a major new idea). It is indeed a matter of facilitating the evolution of new ideas, of new combinations of the factors of production, and generally promoting

innovation. As Youngson (1967) notes, it is here that the idea of external economies meets the Schumpeterian notion of innovation. Innovation is the key to economic development. Infrastructure facilitates investment that promotes innovation. An appreciation of the role of infrastructure in facilitating the emergence of new combinations of factors of production is important. The analysis of infrastructure inevitably includes the study of economic development - which we shall examine next.

3. INFRASTRUCTURE AND INDUSTRIAL TRANSFORMATION

Many regions directly provide capital goods - designed both to supplement and to induce a favorable response from the productive enterprises - in order to take advantage of the beneficial effects of infrastructure. It would appear that the stock of infrastructure has several effects on the level and mix of directly productive activities. First, investments in physical and social overhead capital will increase the efficiency and reduce the prices of production inputs. Not only do costs such as those of material assembly and skilled labor become lower, but increases in the capacity of infrastructure very often lead to an improved quality of service. A six lane limited-access highway not only has a greater capacity than a two-lane road, it is also faster and safer - thereby generating new demands (such as labor and capital).

Figure 1 reflects how traditional theory would conceptualize these effects for a market economy with perfect competition.



MC1 = marginal cost with infrastructure deficiencies
MC2 = marginal cost with adequate infrastructure

Figure 1 Infrastructure Provision and the Efficiency of Production

In a situation of inadequate availability of infrastructure, the firms are confronted with higher marginal costs (MC1) at every level of production, and given the market price of their output produce at a level Q_1. With an improvement in infrastructure services or relocation to areas of higher infrastructure quality (and/or quantity), the marginal cost curve shifts to a lower level (MC2) resulting in total cost savings of abcd for the earlier level of output Q_1, an increase in the output (Q_2-Q_1).

These cost reduction and output expansion effects of infrastructure investments are empirically captured through the formulation and estimations of cost functions and production functions. Since social overhead capital is available to all firms in a region, it

is viewed as entering the production functions of regional firms. However, while available to all, total use must be equal to or less than the physical capacity (e.g. traffic lane capacity, sewage pipe diameter, etc.). Infrastructure is viewed as a stock variable in physical terms.

A number of such empirical assessments are available (CONSAD, 1968; Lakshmanan and Lo, 1970; Mera, 1973; EEC, 1982; Saxonhouse, 1977; Antle, 1983; Fogerty et al., 1983; Wigren, 1985). Essentially, these formulations use a Cobb-Douglas specification of a regional production unit (obtained by aggregating firm level data) with market inputs of labor and capital combined with infrastructure. Output elasticities of various types of infrastructural investments are estimated and interpreted. Further, the complementarities and substitutions between input factors are assessed. In variations of this approach, the efficiency parameters of the production functions are related to the infrastructure investments.

How useful are results derived from a such largely static, equilibrial, short term framework in a context where infrastructure is viewed as providing a new structure of costs and prices? These studies rely on conventional economic concepts and a framework of production functions, output and Allen elasticities of substitution - while they are not necessarily static they have been used in a static way. Yet infrastructure investments are viewed as facilitating economic transformation, in other words as a developmental mechanism. There is a crucial difference between *"being" industrial units* and *"becoming" industrial units*. The key notions are *dynamic* change, and *disequilibrium behavior*. An appropriate analytical framework would integrate production and investment behavior, describe intertemporal patterns of industrial organization, and describe general dynamic response patterns to change and expansion.

The notion of "adjustment costs" provides a vehicle for such a theoretical formulation. Adjustment costs are those associated with increasing the supply of those resources whose production requires significant periods of time: resources such as sophisticated machines, transport capacity, skilled human capital, etc. These adjustment costs, (while initially analyzed as rationalizing distributed lag models) have been generalized by Treadway (1968 pp. 71, 74). Treadway views the firm as producing two types of outputs: (1) the marketable output and (2) relatively fixed, firm specific productive assets (capacity). The cost of the production of the latter is foregone output. The argument here is that the analytical treatment of the role of infrastructure in industrial transformation requires an extension of the traditional cost minimization behavior of the firm to incorporate adjustment costs and endogenized dynamics related to infrastructure accumulation. Exogenously determined, arbitrary and ad hoc treatments of adjustments are inadequate.

A promising effort in this direction has been made in assessing the infrastructure contributions to Indian development (Elhance, 1986; Elhance and Lakshmanan,1986). This study attempts to integrate a variety of theoretical and conceptual strands.

(1) Neoclassical Theory of Production
(2) Flexible Accelerator Literature
(3) Dynamic and endogenized adjustment processes for infrastructure stocks
    Variable (market inputs) and quasi-fixed (infrastructure) inputs enter into the production and cost functions. Market inputs can be varied instantaneously while adjustments to infrastructure stocks involve time lags and a process involving economic choice variables and optimizing behavior. Since infrastructure adjustments are made in the public sector the firm does not face a Treadway type "internal cost of adjustment", hence the use of (Lucas, 1976) "an external cost of adjustment".
(4) Flexible Functional Forms
    Quadratic functional specifications of normalized variable costs and costs of adjustment for infrastructure stocks were estimated econometrically.

The resulting model has been econometrically estimated in order to interpret the role of infrastructure in regional and national industrial production in India (Elhance, 1986). Some key findings of this study reported in Elhance and Lakshmanan (1986) may illustrate the scope of the approach. For example Capital (k) and overall infrastructure (s) are substitutes at the national level; however at the regional level, k and economic infrastructure (s_1) (s_1: transportation, power, irrigation, sewage, water, etc.) are substitutes while k and social infrastructure (s_2) (s_2: education, health, housing, recreation etc.) are complements. Increases in output lead to less than proportional increases in the demand for economic infrastructure. Again, there is some support for the view that production cost reductions result from infrastructure investments. Based on cost considerations, there is also some support for Hansen's (1965) hypothesis of the relative importance of economic and social infrastructure at different stages of development. Finally, estimates of the stock adjustment coefficients suggest that in India economic infrastructure adjusts almost twice as fast as social infrastructure.

The scope of this study could be generalized by relaxing a major assumption that holds that the regional or national "firm" is a simple aggregation of all production units. Yet we know that complete firms (that perform all functions in the production process) react differently to infrastructure than incomplete firms (Wigren, 1985). Firm size (as a qualitative, multidimensional concept) may affect the contributions of infrastructure. Some infrastructure services may be demanded primarily to deal with the consequences of size. Others may be used *to reach* a large size. Consequently, the use of firm level data on production inputs and outputs in a region along with infrastructure stocks may greatly augment our understanding of how firms and industries of different sizes are stimulated by infrastructure investments.

4. SERVICES, INFORMATION ACTIVITIES, AND COMMUNICATION INFRASTRUCTURE

The widely documented transformation occurring in the industrialized countries is the structural shift towards services. The provision of services has displaced the production of goods as the dominant economic activity. In the sixteen OECD countries the proportion of total employment in the service sector had increased from 24% in 1870 to 60% by the early 1980's. The more recent historical expansion of services in six of these countries is evident in Table 1.

Table 1: Long-term Shifts in the Employment Share of Service Sector (1920-1980)

Country	1920	1950	1970	1982
U.S.A.	38.4%	51.7%	61.5%	68.0%
U.K.	43.7%	45.8%	51.9%	62.6%
France	26.9%	33.9%	43.7%	57.2%
W. Germany	27.7%	36.5%	45.9%	51.8%
Italy	18.6%	25.1%	30.4%	50.6%
Japan	22.4%	27.0%	40.0%	

Source: Singelmann (1978), Saxonhouse (1985)

Three hypotheses have been advanced for this relative growth of the service sector (Fuchs, 1968; Stanbeck, et al., 1981). First, given that the services have an income elasticity of demand higher than 1, as real incomes rise, real services per capita grow faster than the proportional rise in income. This factor combined with the complementarity of consumption of goods and services for a range of goods, leads to services accounting for an increasing proportion of income and national employment. Second, as automation and the division of labor increase with technical change, transaction costs and adjustment costs increase. The transaction costs - which represent the costs of bringing the supply and demand sides of increasingly larger and more complex markets together (e.g. uncertainty, asymmetrical distribution of information, risk, coordination, and control) - generate the demand for new dynamic producer services. Adjustment costs incurred in the supply of resources that have long gestation periods (e.g. sophisticated capital goods, highly skilled labor, special organizational resources, etc.) give rise to a variety of human capital and special producer services. Third, given the slower relative growth of productivity in some services, as the economy expands, the share of the services in total employment increases.

The major growth segments of the service sector are:

- •Final services: such as education and health - reflecting the greater need for investments in human capital in an increasingly sophisticated, internationalized area of production of goods and services.
- •Producer services: These result from an increasing division of labor; the growth in size and importance of large corporations; the rising importance of planning, developmental and complex managerial functions; the growth of markets; and changes in the organizational and institutional arrangements of the private sector, and
- •The producer-service-like functions in the public sector: (Stanbeck et al., 1981). These functions reflect the need to protect consumer interests and worker rights, to further equal opportunity; to regulate markets as necessary, to promote national economic interests in a rapidly expanding international system of production, service delivery and exchange.

The growth of these types of services reflects the increasing global nature of business (resulting from dramatic reductions in unit transportation and communication costs). With expanding markets and an increasing division of labor, there is a greater need for coordination, for integration and binding together the increasingly differentiated, specialized parts and functions of the global market system. The greater the number of participants and the finer the division of labor, the more complex the technological process and the greater the range of goods and services an economic system generates, the more "economic intelligence" or "technological and scientific information" the economic process requires. The economic system becomes more information intensive.

Such changes in the information environment accompanying broad structural changes have been noted for over two decades. Machlup's (1962) pioneering analysis of the 'knowledge industries' (including education, research, publishing and broadcasting) suggested that in 1962 over 31 percent of the U.S. labor force was working in such occupations. Bell (1973), using a narrower definition of knowledge workers as only information producers, estimated the core of a knowledge society at 12.2% of U.S. employment in 1963. The concept of the "information sector" as the segment of the workforce that primarily produces, processes, or distributes information goods and services was elaborated by Porat (1975). Porat's broad definition includes not only Machlup's knowledge workers, but also all workers who use information as a productive input irrespective of the industrial sector to which their job belongs. Figure 2, drawn from Katz (1985), reflects some of the different definitions of information workers.

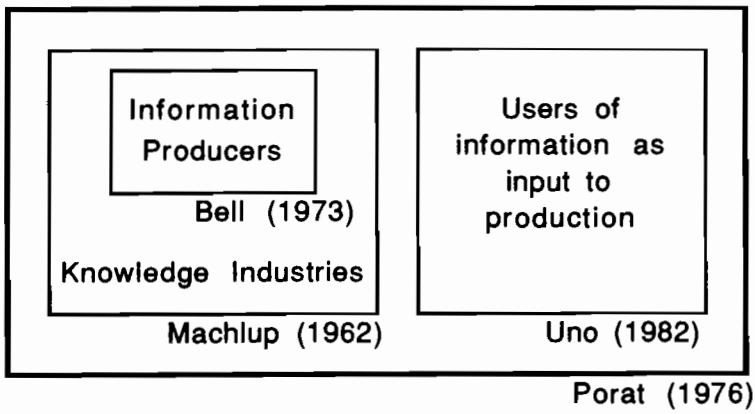


Figure 2 Different Definitions of Information Activities

The longterm trend of the information workforce (Table 2) shows a steady growth in the U.S. and other developed countries (Lamberton, 1982; Katz, 1985). One must note that many information occupations have appeared for the first time in the last few decades and there should be a distinction between the expansion of the information sector from economic growth and the increase in demand for information occupations resulting from increasing division of labor and specialization (Lamberton, 1982).

This new information, communication or knowledge sector concerned with the production of information goods and services and the use of information as a resource includes two components: the information and communication infrastructure and the content and flows of information over that infrastructure.

Table 2: Information Workforce Trends 1840-1980 (% of total workforce)[1]

Country	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
U.S.			5.8	4.8	6.5	12.4	12.8	14.9	17.7	24.5	24.9	30.8	42.0	46.4	46.6
U.K.	4.6	5.1	5.6	6.8	7.9	10.2	12.4	13.3	19.8	20.9	24.4	27.8	33.1	36.6	
Germany												18.3	24.6	30.7	33.2
Australia								8.5	11.5	15.6	16.3	17.0	22.5	27.5	30.2

Source: Lamberton (1982), Katz (1985)

The communication infrastructure is technical in nature and includes telegraph, telephone, radio, television, etc. It provides the highway along which information travels. Recent developments in telecommunications and data processing technologies have reduced structural differences between those two technologies, permitting the connection

[1] The Definition of Information Workforce used is Porat's (1976)

of computers to the telecommunications network and shared service markets and customers. Thus the telecommunications sector is not limited to the traditional transmission of basic voice and message communications, but due to the convergence of market and technology structure, it now provides a variety of voice, text, data, and video transmission services: electronic mail, wide area telephone services (WATS), voice message services, teleconferencing, computerized data transmission, videotex and data base access soft-ware defined 'virtual' networks, environmental control etc. Currently, voice telephone communications and data traffic account for 85% and 10% respectively of the U.S. network. However, the other components are growing very rapidly. Table 3 outlines the rapid growth of the telecommunications services sector in the U.S.

Table 3: The U.S. Telecommunications Sector

	1973	1976	1979	1982	1986
1. Revenues ($ Billion-1982 $)	41.8	51.6	66.5	78.89	97.02
2. Capital Expenditures ($ Billion)	11.8	12.9	20.18	22.53	24.9
3. Cumulative Gross Plant Investment ($ Billion)	94.8	121.6	155.2	202.35	251.0
4. Total Employment (,000)	1,007	961	1070	1100	884
5. Production Workers (,000)	780	730	789	790	659

Source: U.S. Industrial Outlook, 1987

In 1973 and 1983 revenues and capital investment in the telecommunications sector grew at a compounded annual growth rate of 7.9% and 6.6% respectively. Telecommunications infrastructure investments currently account for about one tenth of gross fixed capital formation in the U.S. and Western Europe (Sheferin and Shea, 1987).

As communications activities have grown in importance, interest has also grown in their role in economic growth and development. As Jonscher (1982) has pointed out, there has been an uneasiness in the economic literature about the way in which information viewed as a commodity does not fit into traditional neoclassical economic analysis. Information is not just another commodity, since it has more bearing on both ends and means than other commodities and it is more difficult to handle in a production function when viewed as a resource (Jusswalla and Lamberton, 1982). The structural changes now taking place in various domestic and international economies not only reflect information intensity, but also a great complexity of relationships that are not reflected in the stylized version of the production functions used today.

Furthermore, how does one assess the economic worth of such infrastructure investments in service transformation? The process of producing service outputs is fundamentally different in many aspects from the production of goods. Some scholars have not treated these distinctions as being analytically important or have treated the distinctions as residing solely in the type of output. A survey of the contemporary literature on the service sector persuades us that these differences are important and that a straight-forward transfer of models used to analyze goods production to service production is inappropriate. Before we pursue this matter in the next part of the paper, a

brief summary of the analytically relevant aspects of the ongoing service transformation is in order.

5. THE SERVICE TRANSFORMATION: SOME HIGHLIGHTS

The oft-noted shift to a service economy does not represent a major movement to freestanding services purchased on the market as much as a joint provision of goods and services. Rising real income and increasing diversity of consumer demand lead to greater product differentiation which is followed by joint provision of goods and services. The consumption of services turns out to be strongly complementary to the consumption of goods (Gershuny, 1978; Hirschhorn, 1981; Stanbeck et al., 1981; Daniels, 1985). Furthermore, the reorganization of the production system with the functional differentiation between production and service delivery on the one hand, and central research, administration and control offices on the other, has stimulated a variety of intermediate producer services. We draw attention to three aspects of the service transformation that may help clarify the role of infrastructure in such a transformation.

1. As noted earlier, the growth and diversity of markets, the increasing importance of strategic, developmental and administrative functions and the internationalization of production and service deliveries lead to increases in transaction costs (the costs of bringing different sides of a market together) and in adjustment costs. The variety of producer services that has arisen in response to these costs, is information intensive. The technical advances in computing and telecommunications have vastly increased the potential for cheap information supply. The effect of the vastly increased density, speed, and quality of information flows has been to "transport" a variety of services. The old adage that services cannot be 'stored or transported' is no longer true for many classes of producer and consumer services. At the moment, the larger firms have the resources to take advantage of this telecommunications infrastructure.
2. As Baumol (1967) pointed out, many final (consumer) services with a large labor input, have a 'cost disease', so that longitudinally the price effect acts against the income effect in a manner which reduces the demand for such services. To the degree that some of these services are amenable to some form of 'industrialization', growth is brisk. 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 from the market - usually to the household. Manufactured consumer products (autos, gasoline, T.V. sets, washing machines) are combined with intermediate services (e.g., repair services, T.V. programs), physical infrastructure (roads, broadcast networks, power networks) and unpaid 'informal' (household) labor to produce transportation and entertainment services. Such service innovations that have transformed some consumer services are combinations of *industrial goods, intermediate services, infrastructure, and time.*
3. The distinction between the production of goods (tangible assets) and service outputs noted earlier is clearly drawn in the social indicator literature (Garn, 1973; Garn et al., 1976). Resources - labor, capital energy, materials, land etc. - are combined to produce goods (Figure 3). A production unit is subject to a variety of opportunities to act and constraints on its action. However, the transaction between the producer and consumer will *not* directly affect the level of output. The asset positions of the seller and buyer are altered during sales but the output is not affected by the attitude or behavior of the seller or buyer. This enables the analyst, without serious distortion, to view goods producing units as self-constrained units for analysis of efficiency.

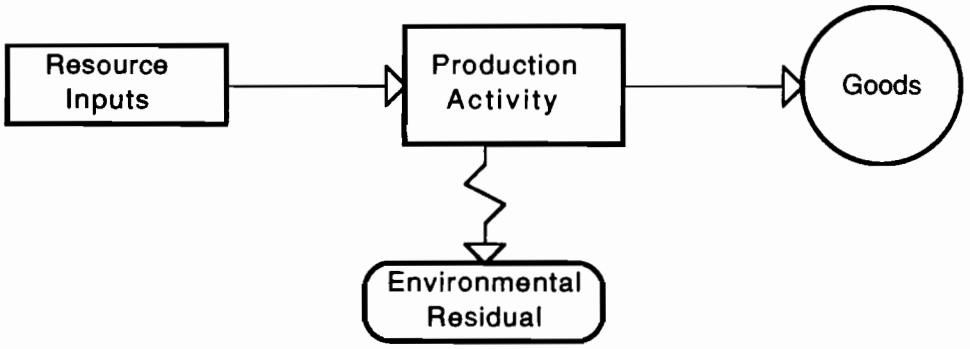


Figure 3 Production of Goods (After Garn et al., 1976)

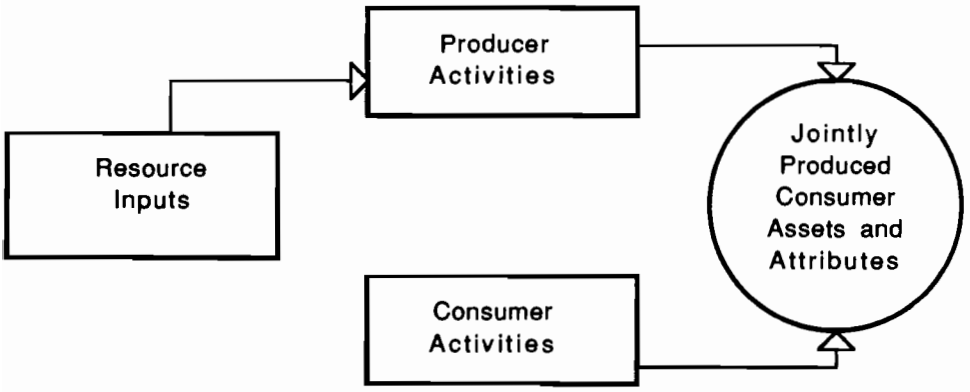


Figure 4 The Production of Service Output (after Garn et al., 1976)

This is not the case with production of services (Figure 4). The consumer of services is part of the process of the production of services. There is *coproduction* of services. In other words, the efficiency of the service production process cannot be attributed solely to the activities of the service provider. The consumer's activities bear upon the efficiency and efficacy of the service.

Further, service activities involve a change in the asset structure of the client or consumer. Indeed, Garn et al. (1976) have argued that there are three major types of services bearing upon such changes. They are:

a) Services in which there is change in the tangible asset structure of both the client and provider, e.g., retail services, repair and maintenance services for durable goods, economic development activities.
b) Services which result in changes in the human capital of the client, e.g., education and training; counseling and health care services.
c) Primarily brokerage activities (e.g., referral progress and real estate brokerage) where the output appears as changes in the personal assets of clients who are providers and clients who are consumers of the service.

6. RELATING INFORMATION INFRASTRUCTURE AND ECONOMIC TRANSFORMATION: SOME CONCEPTUAL ISSUES

What are the analytical implications of these trends (noted above) in technical change, market structure, and organizational form of the rapidly evolving information infrastructure, of the changing perspectives of the production and consumption processes, of the emerging distinction between production of goods and services, and of the complementarity of goods and service consumption? How do we distill all this mass of descriptive-analytical material into an appropriate framework for analyzing the economy in an information-rich age?

Our objective here is the narrower one of relating information infrastructure to economic changes underway. One way to structure the relevant analytical issues raised above is to relate the major economic categories: consumption of goods, consumption of services, production of goods, and production of services (Figure 5).

Quadrant I. The production of goods is a technical problem concerned with handling and processing material resources. The usual formulation of this process is by a highly stylized version that functionally relates a variety of material inputs - capital (K), labor (L), energy (E), and materials (M) - to goods output (X) via the production function:

$$X = X(K,L,E,M). \tag{1}$$

In Section 3 of this paper, we argued that this framework should be broadened to incorporate a vector of infrastructure stock as an additional factor input in order to capture the cost reductions and output expansion effects of infrastructure investments. Or:

$$X = X(K,L,E,M,I). \tag{2}$$

Since infrastructure investments do lead to a new structure of input costs and prices thereby facilitating economic transformation, a short-term static equilibrium framework is not very appropriate. Dynamic change and disequilibrium behavior need to be represented. As noted earlier, the appropriate analytical framework would integrate production and investment behavior and describe general dynamic responses to change and expansion. Consequently, I, can be incorporated as a factor input, where market inputs (K,L,E,M) and I are endogenized in a dynamic framework of adjustment (Elhance and Lakshmanan, 1986). In such a framework, the contribution of infrastructure to goods production can be assessed in a satisfactory manner.

	GOODS	SERVICES
Production	I	IV
Consumption	II	III

	Received Theory	Our Modification	
Quad. I	X = X(K,L,E,M)	X = X(K,L,E,M,I)	
II	U = U(X)		
		Alt I	Alt II
III	U = U(Z,t) Z = Z(X,t) (Becker)	U = U(Z,t) Z = Z(X,t,I)	A theory of consumption of goods, time and locomotion (Lakshmanan and Hua, 1983) (see text)
IV	See section VI of this paper.		

Figure 5 Analytical Frameworks for Production and Consumption

Quadrant II deals with the consumption of goods. Traditional consumer theory assumes that the consumer maximizes utility, a function of goods, subject to a budget constraint. When the system is solved for quantities of goods demanded, demand turns out to be a function of prices of goods and the consumer's income.

$$U = U(X) \tag{3}$$

Quadrant III focuses on the consumption of services. Theoretical developments in this area in the last three decades have been extensions of the traditional theory of goods consumption. The central insight is the notion of a household production and the input of time (t), the ultimate resource in consumption. The seminal contributions are those of Becker (1965), Muth (1966) and Lancaster (1966).

Becker recognizes a two-part process of consumption of services. Utility is derived from time and services (Z) that a household produces by processing a vector of goods X that are purchased from the market and time (t). In other words:

$$\begin{aligned} U &= U(Z,t) \\ Z &= Z(X,t) \end{aligned} \tag{4}$$

In the Lancaster approach, consumption is an activity in which goods, singly or in combination, are inputs generating an output that is a collection of characteristics. Utility or preference orderings are viewed as rankings of characteristics and rankings of goods only indirectly through their characteristics.

$$\begin{aligned} &\text{Maximize } U(Z) \\ &Z = BX \\ &\text{subject to } pX \leq y \\ &Z,X \geq 0 \end{aligned}$$

where

Z = vector of characteristics
X = vector of goods
p = vector of prices
B = matrix of consumption technology
y = income.

In our formulation here, we suggest two alternative ways of modifying the Becker formulation. In the first, we recognize the evolving nature of service consumption. In Section 4 we noted that the generation of services (that confer consumer utility) requires inputs not only of goods and time, but also a variety of infrastructure elements (e.g. roads for driving, broadcast networks for entertainment, telecommunication networks for information processing and retrieval, etc.). Consequently,

$$\begin{aligned} z &= z(x,t,I) \\ u &= u(z,t). \end{aligned}$$

In addition, from our reading of the changing nature of consumption in the new service economy, following Scitovsky (1976) we argue for a distinction between different kinds of time: time associated with unpleasant or boring tasks ("bad time") and time associated with stimulating and rewarding tasks ("good time"). We would suggest that t in both the production and utility functions in (6) could be distinguished in this manner so as to enrich our understanding of service consumption.

Another formulation of the theory is that of household consumption of goods, time, and location (Lakshmanan and Hua, 1983). At the heart of this formulation lies the recognition of the role of locomotion in consumption and locomotion as a household choice variable. Locomotion is undertaken partly to accomplish consumption and partly to enable the household to obtain a lower goods price. Thus travel distance is differentiated into a mandatory component and a discretionary component. Given this role of

locomotion in consumption and the endowed nature of household time, the average household locomotion speed- the linkage between household time and travel distance - becomes a crucial decision variable governing the spatial yield of time. Indeed the notion of average speed as a decision variable to be optimized by the household is a salient feature of our formulaation. Following this, goods prices also become endogenized as household specific prices. Several aspects of consumer behavior can be better understood in the framework of our formulation. For example, a rise in income would shift the consumption pattern for reasons other than just the matter of taste. Some goods will be substituted for more expensive time and extra locomotion speed will be demanded to increase the spatial yield of time. Both substitutions lead to less time to be used in consumption and permit more to be used at work or leisure. Thus decisions on goods consumption, time use, and average household speed become intimately related.

Quadrant IV, concerned with the production of services, has received little analytical attention. We believe there is a need to conceptualize the production of services in the light of our emerging understanding of information activities, and the processes of handling and processing information as inputs into service production.

7. TOWARDS A CONCEPTUALIZATION OF SERVICE PRODUCTION

Product flows and information flows are characteristic of the operation of the economy. Product flows are handled in economic theory by Leontief input-output models and linear programming models. Information flows among economic actors, institutions, and between these institutions and the larger economy on the other hand, resist easy measurement.

Further, information viewed as a commodity does not fit into the conventional or neoclassical framework of economic analysis. An economy is an organization for linking individuals by interaction through technical processes of production and distribution on the one hand, and by communication networks of information flows necessary for coordination and control among the individuals on the other. Consequently, an economy must be viewed as constrained not only by its limited capability for obtaining and processing materials, but also by its limited ability to process the information which is required to organize and coordinate these activities. As Jonscher (1982) points out, it may be necessary to develop a framework in which the role of information and the problem of organization are treated as comparable in importance to the role of conventional input factors and the problem of production.

In recent years, there have been a number of attempts to modify the 'perfect information' assumption of the neoclassical model by making the acquisition and transfer of information necessary. Such attempts to analyze the behavior of market actors began with the work of Stigler (1961) and have since been extended by many others. As Jonscher (1982) points out, these efforts provide an ad hoc approach to information acquisition and use. Commonly, one factor (i.e. the price of a product, the level of risk in an insurance contract, or worker productivity, etc.) is the object of uncertainty, while the rest of the economy operates in the perfect-information mode. Thus, such extensions of the neoclassical theory are unable to capture the central role of information in the economy.

Since communications involve the transfer of information, this has consequences for the analysis of the communications sector. The latter sector consumes a very significant portion of the resources in the economy and operates in a context broader than the framework and variables of economic analyses. Communicating behavior is important in decision processes and decisonmaking, organization development, demand for information goods and services, technology change (and the component elements of incurring transaction costs and adjustment costs). These behaviorally sensitive processes cannot be handled by traditional microeconomic analysis that ignores noneconomic

aspects of the relationships between the information and communications sector and the economic system.

As a consequence, we favour an interdisciplinary systems approach to relating information and communications infrastructure to the broader socioeconomic system. The level of interdependence in the system and the need to draw relevant concepts and methods from several disciplines makes this approach attractive (Streeten, 1982).

An interdisciplinary framework for relating the communications infrastructure to that of the broader social system is proposed in Figure 6. This framework builds on Lakshmanan's (1982) model of a socioeconomic system and the work of Porat (1982), Jusswalla and Lamberton (1982), and Streeten (1982) relating information to national development.

We present a simplified model of the role of communications infrastructure in a socioeconomic system in Figure 6 in two parts. First we identify the four systemic components of a socioeconomic structure: the structure of assets, organizations, incentives, and external links. Second, we describe the production of goods, services and the generation and distribution of income in such a socioeconomic system, as guided by sociopolitical decisions that are aided by an information and communication infrastructure - *in both static and dynamic modes.*

Following Lakshmanan (1982) we initially identify four structural elements of a socioeconomic system:

- Structure of assets (level and distribution of all types of assets - physical, human, status, etc. - that generate income and sociopolitical participation).
- Structure of organizations (the nature of markets and other social-political institutions and the associated manifestations of markets complexity and the division of labor).
- Structure of incentives (that provide a particular matrix of opportunity for individual, collective, or sectoral behavior at one time, and
- Structure of links to the outside world (manifest in the two way flows between the socioeconomic system and the rest of the world in terms of labor, capital, information, physical resources, organizational styles, standards of development, development styles, ideologies, etc.).

The interactions between these four structural elements define the context for growth and development in social systems. Societies differ from one another in their structure of assets, organizations, and incentives, and hence, in their potentials for growth or equity. Some structural elements may change faster over time than others, thereby altering the matrix of opportunities. For our purposes here, structural elements can be viewed as providing a particular context for growth and change.

We shall proceed now to a description of the production of goods and services, which in turn leads to income growth in such a systemic context. Resources are allocated to various production activities by policy decisions made by various social and political organizations and the market place. The production sector also receives critical inputs from two other sources: the organizational environment and the information and communications infrastructure.

At any point in time, the efficiency of production and the flow of goods and services is influenced greatly by the organizational environment as defined by the division of labor and the complexity of the exchange system. The degree of the division of labor and the level of market complexity determine the level of transaction costs. As noted earlier, transaction costs represent the costs of bringing together demand or supply sides of a market and include:

- transportation and communication costs
- measurement costs (to determine the quality, quantity and dimensions of the services exchanged)
- insurance costs (hedge against unforeseen circumstances) and
- enforcement costs.

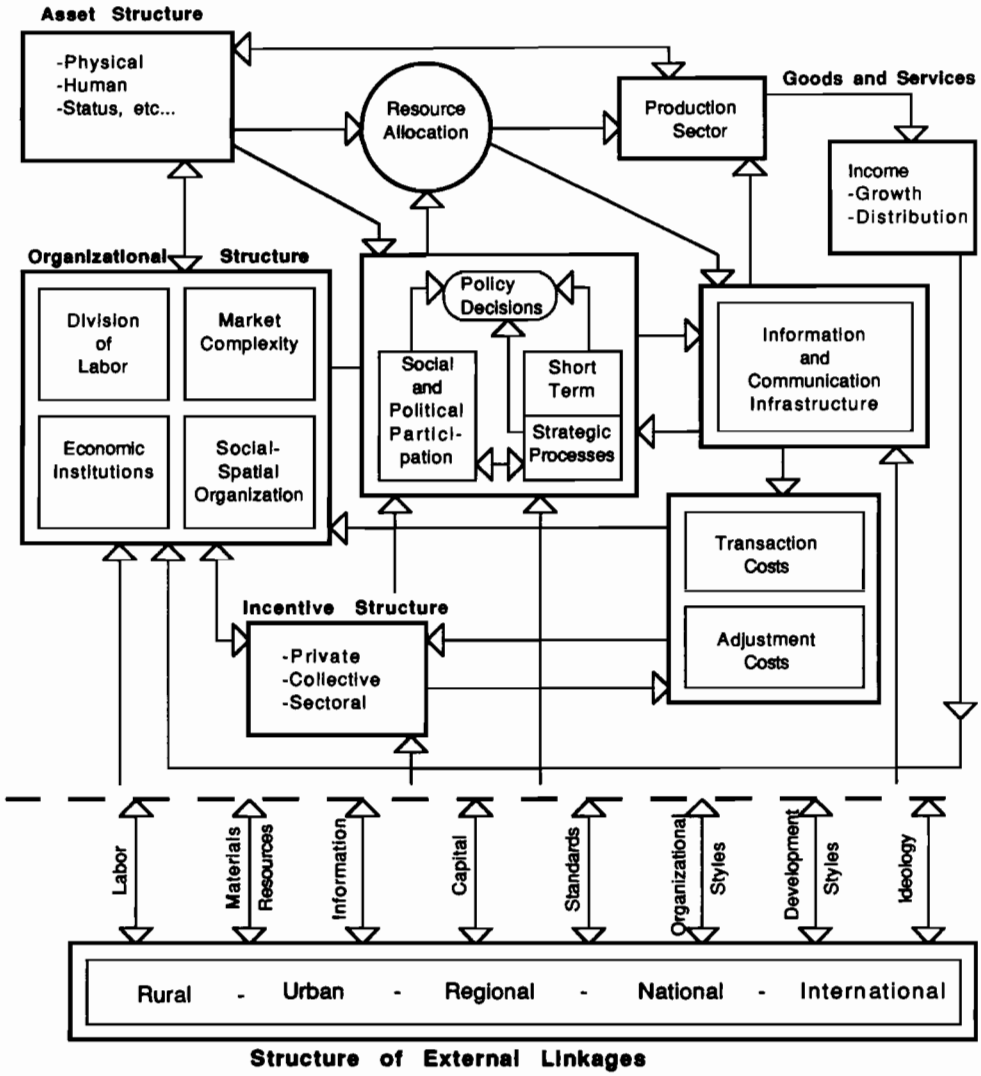


Figure 6 Role of Communication Infrastructure in the Socio-economic System

Such costs relating to spatial and temporal coordination, distribution of information, uncertainty, risk and control are minimized when the relevant information and communication processes operate effectively. Consequently, productive enterprises invest heavily in communications infrastructure that promote cheap information flows in the form of voice, video, data, and text. More generally, as Porat (1982) suggests, the private costs of information and communication and coordination are reduced by a social externality that is generated by social investment in information and communication infrastructure.

As noted above, communication infrastructure includes a variety of information capital (e.g. telephones, satellite communications, integrated digital networks, etc.) and information labor. Such capital reduces the temporal and spatial costs of coordination and over time increases the division of labor and productivity. All of this in turn increases the output of goods and services, income, assets, reinvestment in infrastructure, and institutional complexity.

Thus the information communication infrastructure is viewed as a key sector, receiving resource inputs and imposing transaction costs and making claims on production output as well as providing production inputs. This sector influences the nature and level of social or political participation and the structure of incentives and organizations in society.

However, in the contemporary period of quickening technical change, one must distinguish between states of system maintenence and of system change and development (see Stanbeck et al. (1981)). By affecting the division of labor and market complexity technical change not only changes the structure of transaction costs, but imposes, in addition, a variety of adjustment costs. The latter include the cost of producing more skilled labor, sophisticated capital goods and newer organizations, etc.

In the competitive game of international economic growth, societies that have vigor and adaptability in their social and political institutional structure incur these adjustment costs more effectively and hence speed up their technical change and development. The private and public organizations in such societies adopt a dynamic strategic perspective (in addition to their ongoing system maintenance) and engage in future scanning, goal setting, strategic decision making and programming, that lead in turn to modification of the incentive and organization structures. In such adjustment activities the communication infrastructure also plays a key role.

What we have outlined here is a highly abbreviated version of the model and many more linkages need to be identified. However, it concentrates on the relationship between communication infrastructure and the production of goods and services in a social system both in the short and long run.

Even this is a tall order for theoretical development and our state of knowledge permits no more than an identification of some major elements of such a reconceptualization of economic transformation. We have presented some of the 'building blocks' of theory which deserve further attention from researchers. The central ideas are a reconceptualization of service production, the *organizational, spatial* and *incentive* structure of production and the role of information infrastructure in transforming economic activities in a systemic context.

Some key notions (that need to be developed further) pertaining to *organizational* structure relate to the increasing specialization and the greater need for consequent integration in our society. A major step is to conceptualize the *transaction costs* and *adjustment costs* developing in both private and public sector units in a diverse international economy. The role of knowledge and the interrelationship between the private and public sector need to be analytically treated in this regard.

The notions behind *spatial* structure pertain to a) urbanization and b) the infrastructure facilities and networks. As in the earlier transformation in the 19th century to industrialization, urban structure appears to play a key role in the service transformation. Different types of services appear to be developing in urban areas of different sizes, reflecting the links between new service development and the locational matrix.

Computing and telecommunications infrastructure are behind the information flows that support the international service economy.Quite clearly these ideas need to be elaborated so that many rigorous statements can be made and the various system components linked together. This is a difficult task and it will not be accomplished quickly. But the effort *must* begin.

REFERENCES

Antle, J.M., 1983, "Infrastructure and Aggregate Agricultural Productivity: International Evidence", *Economic Development and Cultural Change*, pp. 609-19.

Baumol, W., 1967, "The Macroeconomics of Unbalanced Growth", *American Economic Review*, vol.

Becker, G:, 1965, "A Theory of Allocation of Time", *Economic Journal*, 75, No. 299, pp. 493-517.

Bell, D., 1973, *The Coming of Post-Industrial Society*, Pergamon Press, New York.

Chatterjee, L., 1986, "Coproduction and Low Income Strategies and Boston", paper presented at the International Housing Research Conference, Gävle, Sweden, June 10-13.

CONSAD, 1968, *The Effects of Public Investment*, Report prepared for the U.S. Dept. of Commerce, Economic Development Administration, Washington,D.C.

Daniels, P.M., 1985, *Service Industries: A Geographical Appraisal*, Methuen, London.

EEC Infrastructure Study Group, 1982, *The Contribution of Infrastructure to Regional Development*, Final Report. June, Brussels.

Elhance, A. P., 1986, Infrastructure-Production System Dynamics in Regional and National Economies: An Econometric Study of the Indian Economy. Ph.D. Dissertation (unpublished), Boston University, Ma.

Elhance, A. P. and T.R. Lakshmanan, 1986, "Infrastructure-Production System Dynamics in the Indian Economy: An Econometric Analysis", Paper presented at European Advanced Summer Institute at Umeå, Sweden, June 16-19.

Fogarty, M.S., R.W. Eberts, and G.A. Garafalo, 1983, "A Model for Measuring the Contribution of Urban Infrastructure to Productivity Growth", Proceedings of Pittsburgh Simulation Conference, University of Pittsburg, April.

Fuchs, V., 1968, *The Services Economy*, National Bureau of Economic Research and Columbia University Press, New York.

Garn, H. A., 1973, "Public Services on the Assembly Line", *Evaluation*, Vol. 1, No. 2.

Garn, H. A., M.J. Flex, M. Springer and J.B. Taylor, 1976, *Models for Indicator Development*, Urban Institute (Paper 1206-17).,Washington, D.C.

Gershuny, J.I., 1978, *After Industrial Society: The Emerging Self Service Economy*; MacMillan, London.

Hansen, N., 1965, "Unbalanced Growth and Regional Development", *Western Economic Journal*, 4:3-14.

Hirschman, A., 1958, *The Strategy of Economic Development*, Yale University Press, New Haven.

Jonscher, C., 1982, "Notes on Communication Theory" in M. Jusswalla and D.M. Lamberton, (eds.), *Communication Economics and Development*, Pergamon Press, Oxford, pp. 60-69.

Jusswalla, M. and D.M. Lamberton, 1982, "Communication Economics and Development: An Economic of Information Perspective" in M. Jusswalla and D.M. Lamberton, (eds.), *Communication Economics and Development*, Pergamon Press, Oxford, pp. 1-15.

Katz, R., 1985, "Measurement and Cross-National Comparisons of the Information Work Force", in *The Information Society*, Vol. 4, no. 4, pp. 231-77.

Lakshmanan, T.R., 1982, "A System Model of Rural Development" in *World Development*, October.
Lakshmanan, T.R. and C.-I. Hua, 1983, "A Temporal Spatial Theory of Consumer Behavior", *Regional Science and Urban Economics*, 13:341-61.
Lakshmanan, T.R. and F. C. Lo, 1970, "A Regional Growth Model for Puerto Rico: Analysis of Municipal Growth Patterns and Public Investment", CONSAD Research Corporation, Pittsburgh.
Lamberton, D.M., 1982, "The Theoretical Implications of Measuring the Communication Sector" in M. Jusswalla and D.M. Lamberton, (eds.), *Communication Economics and Development*, Pergamon Press, Oxford, pp. 36-59.
Lancaster, K. J., 1966, "A New Approach to Consumer Theory", *Journal of Political Economy*, 74:132-57.
Levitt, T., 1976, "The Industrialization of Service", *Harvard Business Review*, September.
Lucas, R., 1976, "Adjustment Costs and the Theory of Supply", *Journal of Political Economy*, August, pp. 331-34.
Machlup, F., 1962, *The Production and Distribution of Knowledge in the U.S.*, Princeton University Press, Princeton, N.J.
Mera, K., 1973, "Regional Production Functions and Social Overhead Capital: An Analysis of the Japanese Case", *Regional Science and Urban Economics*, Vol. 3, No. 2, pp. 157-86.
Muth, R. F., 1966, "Household Production and Consumer Demand Function", *Econometrica*, Vol. 34, 3, July.
Nadiri, M.I., 1970, "International Studies of Factor Inputs and Total Factor Productivity: A Brief Survey", *Review of Income and Wealth*, 18:129-48.
Nurske, R., 1953, *Problems of Capital Formation in Underdeveloped Countries*, Basil Blackwell.
Porat, M., 1976, *The Information Economy*, Ph.D Dissertation submitted to Stanford University, University microfilm, Ann Arbor, Mich.
Porat, M., 1982, "Information, Communication and Divison of Labor" in M. Jusswalla and D.M. Lamberton, (eds.), *Communication Economics and Development*, Pergamon Press, Lamberton, Oxford, pp. 75-80.
Rosenstein-Rodan, P.N., 1943, *The Economic Journal*, June.
Saxonhouse, G.R., 1977, "Productivity Change and Labor Absorption in Japanese Cotton Spinning 1891-1935", *Quarterly Journal of Economics*, 91:195-219.
Scitovsky, T., 1976, *The Joyless Economy*, Oxford University Press, London.
Sheferin, I. H. and T. Shea, 1987, "Telecommunications Services" in *U.S. Industrial Outlook*, Washington, D.C.
Stanbeck, T.M. Jr., P.J. Bearse, T.J. Noyelle and R. A. Karasek, 1981, *Services: The New Economy*, Allanheld, Osmun, Totowa, N.J.
Stigler, G.J., 1961, "The Economics of Information", *Journal of Political Economy*, 69, 3 June, pp. 213-25.
Streeten, P. P., 1982, "The Conflict Between Communication Gaps and Suitability Gaps", in M. Jusswalla and D.M. Lamberton, (eds.), *Communication Economics and Development*, Pergamon Press, Oxford, pp. 16-35.
Treadway, A. B., 1968, "What is Output? Problems of Concept and Measurement", in V. R. Fuchs, (ed.), *Production and Productivity in the Service Industry*, National Bureau of Economic Research, Columbia University Press, New York, pp. 53-83.
Uno, K., 1982, "The Role of Communication in Economic Development: The Japanese Experience" in M. Jusswalla and D.M. Lamberton, (eds.), *Communication Economics and Development*, Pergamon Press, Oxford, pp. 144-58.
Wigren, R., 1985, "Productivity and Infrastructure: An Empirical Study of Swedish Manufacturing Industries and their Dependence on the Regional Production Milieu", *Economic Faces of the Building Sector*, Proceedings from a Swedish American

Workshop on the Role of the Building Sector in the Economy, Swedish Council for Building Research, Umeå, Sweden.
Youngson, A.J., 1967, *Overhead Capital, A Study in Development Economics*, Edinburgh University Press.

ACKNOWLEDGEMENTS

Partial funding for the preparation of this paper provided by the Swedish Building Council at CERUM, University of Umeå, Sweden, is gratefully acknowledged.

Advances in Spatial Theory and Dynamics
A.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 17

On the Dynamics of Regulated Markets, Construction Standards, Energy Standards and Durable Goods: A Cautionary Tale

J.M. Quigley and P. Varaiya

1. INTRODUCTION

Housing is different from many other commodities because of its long life and because it is quite costly to transform the housing stock once it has been put into place. The consequences of a decision about the design or configuration of a dwelling are felt by investors and consumers for forty years or more in the future.

The construction of housing is also affected by the important role of regulatory authorities in societies of very different market orientations. Many of the regulations imposed are designed to control market externalities, to protect consumers and producers from the consequences of their own ignorance, or otherwise to regulate public health and safety.

Certain other construction regulations are intended to enforce cost minimization or to promote efficiency in the market. Energy standards are a conspicuous example. These standards include various regulations specifying the amount of insulation, window glazing, and other aspects of the dwelling which affect its thermal properties. In contrast to regulations based upon externalities or information costs, these regulations can be evaluated by a straightforward benefit-cost test. Such regulations promote efficiency (or at least do not interfere with cost minimization) as long as the discounted savings arising from reduced energy utilization (evaluated at the social cost of energy) exceed the increased capital outlay in construction.

As noted, however, the appropriate time horizon for these calculations is quite long indeed, and the path of exogenous prices is subject to great uncertainty. Despite this, the design of such regulations and the existing economic analyses of the standards have been conducted in a static (or at least stationary) framework.

This paper considers these problems from a dynamic perspective. We analyze the problem of a consumer-investor who must choose a level of initial investment to produce a flow of housing services over some time horizon. Given the initial investment, the housing service flows obtainable from increased variable inputs are deduced. We characterize the optimal path associated with the solution to this problem.

We then particularize the problem by choosing explicit functional forms and parameters based upon related empirical research on the effects of standards in the housing market. We also consider how the motion of the system is affected by regulatory intervention which specifies or restricts the level of initial capital investment. The results presented

here extend existing analyses of regulation in the housebuilding and residential construction industries.[1]

Finally, we point out some serious limitations of this study. Within the context of the model, these have to do with the inadequate treatment of uncertainty about future prices, and the difference in individual and social treatment of time horizons. Beyond the specifics of this model, the limitations reflect a narrowly defined notion of the objectives of standards in the housing market.

2. THE MODEL

A. The Dynamic Optimization Problem

Consider the dynamic problem of the consumer who must choose at $t = 0$ the amount of real estate investment (e.g., insulation, glazing, etc.) to maximize utility over the relevant time horizon T. Let

$$U(H_t, X_t) \tag{1}$$

be the consumer's utility function for housing services (H_t) and other goods (X_t) at time t.

Assume housing services are produced according to the relationship

$$H_t = f(R, V_t) \tag{2}$$

where the level of real estate, R, is chosen initially, and operating inputs V_t are adjusted at each point in time t. Let μ_t be the unit price of operating inputs. For convenience (and without loss of generality) assume that the unit prices of real estate and other goods are each equal to unity.

Denote the consumer's income at time t by Y_t and stock of savings by W_t. At an interest rate of $\bar{\delta}$, the rate of change of savings is

$$\dot{W}_t = \bar{\delta}W_t + [Y_t - X_t - \mu_t V_t] . \tag{3}$$

The consumer must make real estate investment at $t = 0$ when wealth is

$$W(0) = 0. \tag{4a}$$

Discounted savings over time horizon T must defray the initial investment

$$W(T) = e^{\bar{\delta}T}R. \tag{4b}$$

The consumer's problem is thus to maximize utility discounted at rate δ

$$\max \int_0^T U(H_t, X_t)e^{-\delta t}\,dt = \max \int_0^T U(f[W(T)e^{-\bar{\delta}T}, V_t], X_t)e^{-\delta t}\,dt. \tag{5}$$

[1] See Colwell and Kau (1982) or Quigley (1982) for an analysis of the effects of the regulatory environment on housebuilders and final consumers.

Utility maximization is subject to the constraints imposed by equations (2), (3), and (4). Equations (3) and (4) can be combined into a single budget constraint:

$$\tilde{Y} = \int_0^T Y_t e^{-\bar{\delta}t} dt = R + \int_0^T \mu_t V_t e^{-\bar{\delta}t} dt + \int_0^T X_t e^{-\bar{\delta}t} dt , \tag{6}$$

and the income stream Y_t can be represented by an initial endowment, $\tilde{Y}$. To examine the properties of dynamic equilibrium for the consumer, consider the Hamiltonian

$$\theta(W,X,V,R,\lambda,T) = U(f[R,V_t],X_t)e^{-\delta t} + \lambda_t [\bar{\delta}W_t + (Y_t - X_t - \mu_t V_t)] . \tag{7}$$

At the optimum values λ^*_t, W^*_t, and R^*, the values of X^*_t and V^*_t must be chosen to maximize the Hamiltonian θ. At the optimum

$$\frac{\partial\theta}{\partial X} = \frac{\partial U}{\partial X} e^{-\delta t} - \lambda^*_t = 0 \tag{8}$$

and

$$\frac{\partial\theta}{\partial V} = \frac{\partial U}{\partial H}\frac{\partial H}{\partial V} e^{-\delta t} - \lambda^*_t \mu_t = 0. \tag{9}$$

The adjoint equation is

$$\frac{\partial\lambda^*}{\partial t} = \frac{-\partial\theta}{\partial W^*} = -\bar{\delta}\lambda^*_t \quad , \quad \text{or} \tag{10}$$

$$\lambda^*_t = e^{-\bar{\delta}t}\lambda^*_0 . \tag{11}$$

Substitution into equations (8) and (9) yields

$$\frac{\partial U}{\partial X} = e^{(\delta - \bar{\delta})t}\lambda^*_0 \tag{12}$$

and

$$\frac{\partial U/\partial X}{(\partial U/\partial H)(\partial H/\partial V)} = \frac{1}{\mu_t} . \tag{13}$$

Finally, substituting for H_t in equation (5) and differentiating with respect to W_T yields

$$\lambda^*_T = \frac{\partial}{\partial W_T}\int_0^T U(f[e^{-\bar{\delta}T} W_T, V_t], X_t)e^{-\delta t} dt \tag{14}$$

or, using (11) and (12)

$$\frac{\partial U}{\partial X} = \int_0^T \frac{\partial U(H^*_t, X^*_t)}{\partial H} \frac{\partial H(R^*, V^*_t)}{\partial R} e^{-\delta t}\, dt\,. \tag{15}$$

Using (12) and (13), this expression can be simplified to

$$e^{(\bar{\delta}-\delta)t} \frac{\partial U}{\partial X} = \int_0^T \frac{\partial U}{\partial X} \mu_t \frac{[\partial H(R^*, V_t^*)/\partial R]}{[\partial H(R^*, V_t^*)/\partial V]} e^{-\delta t}\, dt\,. \tag{16}$$

To simplify these calculations, we assume $\bar{\delta} = \delta$; then $\partial U/\partial X$ is constant, and we get

$$1 = \int_0^T \mu_t \frac{[\partial H(R^*, V_t^*)/\partial R]}{[\partial H(R^*, V_t^*)/\partial V]} e^{-\delta t}\, dt\,. \tag{17}$$

Equations (11), (12), (13) and (17) represent the equations of motion of the system. For a given income stream Y_t, interest and discount rate δ, and relative price stream μ_t, the dynamic pattern of consumption of operating inputs V_t^*, other goods X_t^*, and the initial investment in real estate R^* are determined, along with the adjoint variable λ_t^*.

B. The Case of Separable Utility

The solution to this dynamic system can be further simplified if the utility function is separable in its arguments. Suppose, δ and $\tilde{Y}$ are given and that the utility function is

$$U = k(H_t) + g(X_t)\,. \tag{1'}$$

Under these circumstances, the dynamic system can be solved by trial and error. First, pick a trial value of the adjoint variable λ^*_0. Solve for X^*_t from (12)

$$\frac{\partial U}{\partial X_t} = \frac{\partial g}{\partial X_t} = \lambda^*_0\,. \tag{12'}$$

Then solve for V^*_t as a function of the as yet unknown R^*, from (13)

$$\frac{\partial U}{\partial H}\frac{\partial H}{\partial V} = \frac{\partial k}{\partial H}\frac{\partial H}{\partial V} = \lambda^*_0 \mu_t\,. \tag{13'}$$

Then calculate R^* such that equation (17) is satisfied. Finally, check to see that the budget constraint is satisfied:

$$\tilde{Y} = R^* + \int_0^T \mu_t V^*_t e^{-\delta t}\, dt + \int_0^T X^*_t e^{-\delta t}\, dt\,. \tag{6'}$$

If the right hand side (RHS) of (6') exceeds the endowment $\tilde{Y}$, increase the trial value of λ^*_0. If the RHS of (6') is less than the endowment, decrease λ^*_0.

For the separable case, as long as equation (13') can be solved for V^*_t (as a function of R*) in closed form, then equation (16) can be solved for R*, by numerical methods if necessary. If utility is not separable in H and X, then the model may be much more difficult to solve, since (12') and (13') must be solved simultaneously for X^*_t and V^*_t, given a trial value for λ^*_0.

3. SOME IMPLICATIONS

According to the model presented in Section 2, the investor-consumer chooses the amount of real estate to purchase initially, assuming technical efficiency in production and knowledge of the time path of the price of operating inputs over the relevant planning horizon. The consumer makes this choice to maximize the present value of lifetime utility.

In this section, we use this model to analyze the imposition of standards which specify the amount of real estate R to be used in the production of housing. The analysis is highly stylized, to be sure, but we utilize empirical functions from a real housing market characterized by the recent imposition of energy standards. These standards are intended to increase the real estate component of housing and to economize on the operating costs - principally energy - associated with the flows of housing services enjoyed in final consumption.

We assume a separable utility function of the following form

$$U_t = \alpha H_t^{\beta} + X_t^{\varepsilon}, \tag{1''}$$

where α, β, and ε are parameters.

We also assume that housing services are produced from real estate and operating inputs according to a Cobb-Douglas technology

$$H_t = AR^{\gamma_1} V_t^{\gamma_2} = R^{\upsilon} V_t^{1-\upsilon}. \tag{2''}$$

By suitable choice of units of measurement this process can be represented by the single parameter υ.

The parameters of these relations, α, β, υ, ε are estimated from information about newly constructed single family dwellings and their occupants. These observations were obtained from U.S. government mortgage applications (under the Federal Housing Administration) in California during the period 1974-1978, before energy standards were introduced.

The production function is estimated from information on housing expenditures ($P_H H$), operating expenditures ($P_V V$), and real estate expenditures ($P_R R$) for this sample of dwellings. The parameters of the utility function are estimated from the first order conditions for utility maximization,[2] that is from a regression of the price of housing (i.e., the marginal cost of housing from the production function) on the quantities of housing and "other goods" (i.e., income minus housing expenditures) available for each household. A detailed description of the underlying data is available elsewhere.[3]

---

[2] From (1") and the budget constraint

$$\frac{\partial U/\partial H}{\partial U/\partial X} = \partial P_H/\partial H = P_{H'} = \frac{\alpha\beta H^{\beta-1}}{\varepsilon X^{\varepsilon-1}}$$

[3] See Quigley (1985) for a discussion of the underlying data and methdology. For present purposes it is important to note that the estimation of the production function assumes technical efficiency in the

Table 1 presents regression estimates of the two equations. In the regression of housing expenditures on real estate expenditures, constrained so that the intercept is zero, about three quarters of the variance is explained and the t-ratio is quite large by any standard. The estimate of υ, exp (-.1352), is 0.874. The second equation explains about two thirds of the variance in the dependent variable, the log of the marginal price of housing, and the three coefficients are highly significant. The results imply values of α, β, and ε of 5.343, 0.974, and 0.910 respectively.

Table 1: Estimates of Utility and Production Function Parameters: 7378 California Households and Single Family Dwellings (t-ratios in parentheses)

a. Regression Estimates:

1. Production

$$\log P_H H = 0 + [\log \upsilon] \log P_R R$$
$$\log P_H H = 0 - 0.1352 \log P_R R \quad (17.22)$$
$$R^2 = .766$$

2. Utility

$$\log P'_H = [\log \alpha\beta/\varepsilon] + [\beta - 1] \log H + [1 - \varepsilon] \log X$$
$$\log P'_H = 1.7442 - 0.0261 \log H + 0.0901 \log X \quad (11.14)\ (16.73)\ (3.39)$$
$$R^2 = .677$$

b. Implied Functions:

1. Production

$$H = R^{\upsilon} V^{1-\upsilon}$$
$$H = R^{.874} V^{.126}$$

2. Utility

$$U = \alpha H^{\beta} + X^{\varepsilon}$$
$$U = 5.343\, H^{.974} + X^{.910}$$

With these parameters and the assumed functional forms, the dynamic system may be expressed more simply. Normalize the initial price of operating inputs, $\mu(0) = 1$, and assume that the expected price path for operating inputs is exponential

$$\mu_t = e^{\omega t}. \tag{18}$$

Finally, choose the units in which the initial endowment is given so that $\lambda^*_0 = 1$. Under these circumstances the system simplifies, after some manipulation, to:

$$\frac{\partial U}{\partial X} = .91\, X_t^{.09} = \lambda^*_0 \tag{19}$$

$$\frac{\partial U}{\partial H}\frac{\partial H}{\partial V} = .66 R^{.85} V_t^{-.88} = \lambda^*_0 e^{\omega t} \tag{20}$$

$$1 = \int_0^T 6.93 \frac{V}{R} e^{(\omega-\delta)t} dt = \frac{11.11 \lambda^{*-1.14}_0}{R(.14\,\omega + \delta)} [1 - e^{-(.14\,\omega + \delta)T}]. \tag{21}$$

The equilibrium consumption path is thus

---

production of housing. The estimation of the first order conditions for utility maximization exploits individual variation in housing prices arising from variations in land prices and geographical variation in energy and captial costs.

$$X^*_t = 2.85 \tag{22}$$

$$V^*_t = 6.68\left(\frac{1 - e^{-(.14\,\omega + \delta)T}}{(.14\,\omega + \delta)}\right).97\, e^{-1.14\,\omega t} \tag{23}$$

$$R^* = 11.10\left(\frac{1 - e^{-(.14\,\omega + \delta)T}}{(.14\,\omega + \delta)}\right) \tag{24}$$

$$H^*_t = 7.74\left(\frac{1 - e^{-(.14\,\omega + \delta)T}}{(.14\,\omega + \delta)}\right).87\, e^{-0.14\,\omega t}. \tag{25}$$

Given the planning horizon, T, the interest rate, δ, the expected rate of price increase for operating inputs, ω, and the endowment $\tilde{Y}$, the initial investment in real estate R^* and the stream of operating inputs V^*_t are chosen. These choices define the stream of housing H^*_t and other goods X^*_t to maximize lifetime utility.

Consider the problem instead from the perspective of the public regulator who chooses an energy standard for newly constructed dwellings. Based upon expectations about the rate of price increase $\omega\#$, the regulator chooses a standard, R#. If ω equals $\omega\#$, then the minimum standard is the same level as would be chosen in the absence of regulation. If $\omega\#$ is less than ω, then the minimum standard is not binding; investors and consumers choose a level of initial real estate investment that exceeds the mandated minimum. If, however, $\omega\#$ exceeds ω, then the consumer is forced to consume more "conservation" than is warranted by expected price increases over the planning horizon. Of course, if the regulator's forecast of price increases is more accurate than the private market's, then consumers are better off under the regulatory regime. If the consumer's forecast is more accurate than the regulator's, then the dynamic losses from regulation may be substantial.

Table 2 uses the theoretical model and the parameter estimates from the California housing market to investigate the dynamic efficiency of regulation under various assumptions about increases in the price of operating inputs. For a given planning horizon (T = 40 years) and interest rate, the optimal real estate investment, as well as the resulting discounted utility level, can be computed for given rates of energy price increase. Given an arbitrary level of real estate investment, it is also possible to compute the discounted utility level for the consumer, for various rates of energy price increase. Thus, for example, for the level of real estate investment which would be optimal with a 10 percent rate of energy price increase, it is possible to compute the consumer's level of well being if energy prices increase by only 4 percent.

Each entry in the table represents the discounted utility lost by overly restrictive standards. The base is the discounted utility computed from the optimal real estate investment associated with a particular rate of energy price increase. For the same rate of energy price increase, the discounted utility is also computed using a mandated level of real estate investment which assumes higher rates of price increase. The table reports the percentage difference in these levels of well being. The entries in the table thus indicate the dynamic inefficiency of regulatory standards which are too high relative to the optimally chosen level of real estate investment.

As the table indicates, the utility losses associated with overly stringent construction standards may be quite large indeed. Consumer welfare may be reduced by 15-25 percent if construction standards are based upon expectations of high rates of energy price increase relative to observed rates of increase.

Table 2: Consumer Losses from Overly Stringent Construction and Energy Standards

	Rate of Increase Assumed by Standard				
	5%	8%	10%	12%	15%
Interest rate, δ					
5%					
Actual Rate of Increase, ω					
2%	11%	19%	28%	33%	36%
4	8	11	13	29	34
6	0	6	11	20	27
8	0	0	5	13	19
10	0	0	0	4	11
Interest Rate, δ					
10%					
Actual Rate of Increase, ω					
2%	12%	21%	29%	35%	38%
4	9	13	17	30	35
6	0	8	14	24	29
8	0	0	7	16	22
10	0	0	0	7	14

Entries represent discounted lifetime utility lost from overly stringent standards, in percent: $100(U_1 - U_2)/U_1$. U_1 is discounted utility at given values of δ and ω in the absence of regulation. U_2 is discounted utility at the same values, given regulations which set R equal to the optimal level or each column entry. For example, at a 5% interest rate and a 2% rate of price increase, the total loss if efficient standards are set in anticipation of a 5% rate of price increase is 11 percent. Utility is computed from the solution to equations (6), (22), (23), (24) and (25).

4. CONCLUSIONS

This paper analyzes a model of housing formulated to make explicit the choice between increased initial investments in insulation and other energy saving capital, on the one hand, and increased operating costs on the other hand. Crucial to the optimum level of investment is the expected increase in future energy prices.

The model is solved for numerical values using functional forms and parameter estimates for operating expenses and production relationships derived from a sample of newly constructed dwellings and their occupants in the California housing market during the period 1974-1978. Using these values, the model analyzes the dynamic efficiency associated with mandated energy standards, and computes the losses associated with regulation based upon overly pessimistic assumptions about the future course of energy prices.

According to the parameters and the solutions to the model, consumer well being could suffer by 15-30 percent from overly restrictive standards.

It is, of course, always dangerous to draw firm conclusions from overly simple models. The specific results and their application to a particular market depend upon a variety of simplifying assumptions and estimated parameters. Nevertheless, it is worth noting that the mandatory energy standards adopted by the California Energy Commission (1981) assume a real compound increase in energy prices of almost ten percent. The course of energy prices in the United States has been quite stable during the recent past; increases have averaged less than 0.5 percent a year during the past 7 years. If these trends continue, the losses from these standards may be quite large indeed.

However, one should not draw such a conclusion without appreciating the major limitations of this analysis. There are two sets of limitations: the first accepts the context of the model; the second goes beyond it. The crucial features about housing are extreme durability, the possibility of substituting between increased investment for insulation, etc. and increased operating expenses for energy, and the fact that the optimum investment level depends on future energy prices.

The model analyzed in this paper clearly indicates how investment decisions based upon unrealized forecasts can lead *ex post* to substantial losses. It would, however, be incorrect to conclude on this basis alone that the particular California energy standards adopted in 1981 led to losses due to the pessimistic forecast of price increases. To substantiate such a conclusion one would have to show that the forecast was unduly pessimistic given the information available at the time it was made.

There is a rather different reason for higher mandatory construction standards arising if the time horizon T of the consumer-investor is too small. From equation (24) we observe that the smaller is T the smaller will be the optimal initial investment R*. If it is costly to change R* once a building has been completed, then the initial value of R* will be less than the social optimum.

The second set of limitations goes beyond the context of the model. Energy conservation measures (including building standards) may be based on considerations that go well beyond narrow standards of economic efficiency of the kind considered here. They may be based on concerns of ecology, national security, etc. Even granted, however, that these broader concerns lead to social decisions to conserve energy, the kinds of calculations suggested in this paper are important in deciding how much conservation should be encouraged in different activities.

ACKNOWLEDGEMENTS

This paper was originally supported by the University-Wide Energy Research Group, University of California, Berkeley. Additional research support has been provided by the Center for Real Estate and Urban Economics. We are grateful for an extraordinarily helpful review by an anonymous referee.

REFERENCES

California Energy Commission, August 1981, Energy Conservation Standards for New Residential Buildings, processed.

Colwell, P.F. and J.B. Kau, 1982, "The Economics of Building Codes and Standards", in M.B. Johnson, (ed.), *Resolving the Housing Crisis*, Ballinger Publishing Company, Cambridge.

Quigley, J.M., 1982, "Residential Construction and Public Policy: A Progress Report" in R.R. Nelson , (ed.), *Government and Technical Progress*. Pergamon Press, New York.

Quigley, J.M., 1985, "The Production of Housing Services and the Derived Demand for Residential Energy", *The Rand Journal of Economics*, vol. 14, 4:555-567.

Advances in Spatial Theory and Dynamics
Å.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 18

Dynamic Energy Complex Analysis for Metropolitan Regions

H-H. Rogner

1. INTRODUCTION

Contemporary energy planning for large metropolitan areas calls for a comprehensive system's approach encompassing both the energy production and consumption aspects of metropolitan energy systems. Until recently, metropolitan energy planning concentrated primarily on planning for the timely availability of adequate quantities of energy supplies, a process which traditionally has been the occupation of utilities. Energy demand was considered to be closely linked to economic activity. Since it has been desirable for the latter to grow steadily, so have the projected energy requirements. In this respect, comprehensive metropolitan energy planning did not present itself as an area for inter-institutional or interdisciplinary involvement.

Although not directly the focus of institutional planning activities, energy-related planning occurred indirectly within the traditional urban planning framework. The activities of planning departments, such as housing and residential developments, urban transportation systems, commerce, health and medical care, etc., all have an energy component in one way or another and to varying degrees. As long as energy was a relatively inexpensive commodity, urban planners tended to treat it as a residual.

The high energy cost level of the last decade has unveiled the recognition that urban planning directly affects energy consumption. In turn, the adverse aspects of energy production (conversion) and its use in densely populated areas have had an increasing impact on the thinking processes of urban planners. Consequently, there has been a growing demand for comprehensive metropolitan energy planning. Given the complex interaction between energy and all other urban activities, it seems obvious that sector-specific planning with its dispersed responsibilities does not adequately meet the needs and challenges of tomorrow's metropolitan areas. Thus, efficient and effective metropolitan energy planning must encompass all other aspects of urban planning. In turn, non-energy urban planning processes have to account for the energy relevance of their eventual implementation.

From the above, we may conclude that integrated or synergistic metropolitan planning supported by system analysis has become the crucial concept in the development and design of future metropolitan structures. This is to say that regional planners of all relevant departments, together with scientists with expertise in a wide range of interdisciplinary areas, have to be jointly involved in finding a solution rather than competing against each other, which has often been the case to date.

The Stockholm County energy study was an attempt to meet the requirements of an integrated energy study. Although this paper focuses primarily on the energy system's aspects, the energy deliberations should be seen within the overall context of a comprehensive metropolitan planning exercise. The results of various studies on the creation of new residential areas, housing development, demographics and migration issues, industrial development, urban transportation, service requirements, etc. became integral parts of the urban energy demand and supply analysis.

The point of departure for the Stockholm study was marked by the turmoil caused by the second price rise in the international oil market (1979/1980). Apart from the looming economic impacts of oil prices approaching the 40$/barrel level, Sweden, as many other economies, also began to feel the consequences of the rigid energy-economy correlation of the past. The abruptly changed energy market conditions of the early 1970s have caused unexpected responses of all economic agents. Though slow initially, the effects of high energy price levels eventually translated into energy demand reductions. For an economic sector like energy, where construction periods for capacity expansion run up to ten years or more, any disturbance of anticipated demand growth rates can have disastrous effects. For example, by the end of the 1970s the anticipated growth in the Swedish electricity production capacity exceeded actual demand. However, electricity production capacities were installed so as to meet forecasted demand comfortably. By the beginning of the 1980s significant surplus capacity and the threatening risk of huge economic losses forced the utility sector to search for alternatives to utilize these capacities.

Similar examples can be found in other energy sectors, e.g. the underutilization of large parts of the world's oil refining capacities. To a certain extent, such mismatches can be attributed to the sluggish performance of the economies worldwide. But there are other factors, long term by nature and not reversible as soon as the world economy recovers or energy prices decline. These include energy conservation, interfuel substitution, changes in consumption behavior and structural economic change.

It was the problem of surplus production capacity and, at the same time, the uncertainties regarding the future evolution of those factors causing the decoupling of energy use from economic activity which triggered the Stockholm metropolitan energy study. At the outset, a number of generally applicable guidelines for the performance of such a study were determined:

- Energy planning for metropolitan regions requires the simultaneous analysis of all stages of the energy chain, i.e. from resource extraction, energy conservation and/or energy transport, transmission, distribution, storage, energy end-use conversion to the determination of energy services.
- The determination of energy services, i.e. the areas heated in privates homes and office buildings, the quantities of low and high temperature steam for commercial and industrial processes or the ton-kilometers in freight and passenger transportation, are the crucial factors determining energy demand.
- The translation of the energy service requirements into energy demand must be based on information about the age structure of the building stock, the existing and anticipated insulation standards, the efficiencies of energy consuming equipment in industries, in the service and transport sectors as well as in private households.
- The longevity of urban structures poses a problem in itself. Survival functions of the existing infrastructure will have to be specified in order to define market penetration rates for new and more energy efficient equipment. By the same token, depreciation functions will be used to determine the speed of interfuel substitution or the substitution of other factors of production for energy within industries.
- Factors limiting the continued utilization of various technologies or the introduction of new ones may originate from environmental considerations. For example, the decentralized combustion of coal in private homes or small block heating plants - as

part of an oil substituting strategy - certainly conflict with air quality standards in densely populated metropolitan areas.

- Of particular importance for urban planning is the energy distribution system, i.e. the link between the locations of energy consumption and the sites of the central conversion plants. Electricity grids, district heat systems or gas pipeline networks are deeply embedded in urban infrastructure. The interdependence of grids and infrastructure must be carefully considered. The economics of grids are usually highly dependent on energy demand densities and must be weighed against grid independent energy alternatives in low energy consumption areas.
- The appropriate central conversion technologies must be determined and the siting question addressed. Environmental consequences must be taken into consideration as well as the provision of sufficient back-up capacities in cases of unscheduled down times of large central conversion plants.
- An essential objective in the design of modern energy systems concerns the question of robustness or resilience against external shocks. Energy systems are capital intensive and therefore should not collapse upon sudden changes in the world energy market condition. Hence, a certain degree of flexibility has to be included. This points to the conflict of achieving flexibility, economic optimality and high energy efficiency simultaneously.

Today, energy planning for metropolitan regions often means the adaptation of high energy consumption densities - areas which in the past have been primarily supplied with grid-independent energy forms - to new, advanced and most likely grid-dependent energy supplies. This transition is confined by infrastructures which are often centuries old, a constraint which poses a particular challenge to the urban planner. But high energy consumption densities offer an opportunity for tackling the energy question where things matter:

1. Even small efficiency improvements along the entire energy chain result in considerable reductions in overall energy consumption (and pollution emissions);
2. The embedding of modern energy systems in existing infrastructures favors the concept of synergism; and
3. The simultaneous analysis and evaluation of all stages of the urban energy system together with the inclusion of the insights rendered from the traditional urban planning departments help avoid the implementation of conflicting urban development policies.[1]
4. The energy demand in the metropolitan areas of the northern hemisphere to a large extent concerns the demand for low and medium temperature heat for space and water heating. This opens the opportunity for the introduction of a synergetic system design that incorporates the heating technologies in private homes and commercial buildings as quasi-cooling devices of the large-scale energy production and conversion plants. Such a configuration would improve the overall energy efficiency considerably. Similar opportunities can hardly be found in rural areas which often have large distances between villages. Although technically within reach, the economics involved lack the scale economies offered by areas with high energy densities.

2. THE CONCEPT OF NOVEL ENERGY SYSTEMS

The request for robustness of modern energy systems relates directly to the future prospects for the global energy situation. A comprehensive outlook of the future global

[1]For example the introduction of stringent energy conservation measures may reduce the energy consumption density to a level which makes the erection of grid dependent energy supply alternatives economically infeasible.

energy resource situation was reported in the IIASA study *Energy in a Finite World* (Häfele, 1981). One of the study's main objectives concerned the transition from the current global energy system, based on depletable fossil fuels, to a sustainable system based on nondepletable fuels. Within this analysis, the study also reexamined the depletion phenomena of fossil fuels as propagated by the Club of Rome (Meadows et al., 1972) in the early 1970s. The IIASA analyses showed that the depletion problem will not take place nor will the transition towards a sustainable energy system be achieved during the next 50 years. The transition process will begin gradually at the end of this century, and its completion will not occur until the end of the 21st century. The study concluded that although the transition towards a sustainable energy system will not be accomplished within the next five decades, there will be a transition of a different kind. In contrast to the anticipated shift away from fossil fuels, the world will utilize fossil fuels at an increasing rate. This fossil-to-fossil transition is characterized by a growing consumption of low grade fossil fuels (low hydrogen-to-carbon ratio), but in a different technological frame from that in the past.

The prospects of utilizing even dirtier fuels than today with today's energy technologies mean at least a proportional increase in environmentally harmful quantities of hazardous emissions. An increase in pollution emissions is in conflict with current efforts to minimize the burden to the environment in many regions. An acceptable resolution of this conflict could be an energy conversion system that is characterized by quasi zero emissions.

Concepts of energy systems with zero emissions have been advanced by Häfele et al., (1982), Sassin (1982), Rogner (1982), and Häfele et al. (1986). Zero emission energy systems imply a systems configuration where the purification, waste handling, quality and pollution control is shifted up-front (i.e. from the consumer to the producer of secondary energy). The secondary energy leaving this front-end conversion stage will be of high quality and essentially free of impurities. Examples are electricity, district heat or town gas systems. All these systems have common features - they are grid dependent and subject to significant economies of scale. However, these examples are not convincing as prototypes for future zero emission systems, since the conversion stations for these forms of energy are at present the main sources of CO_2, SO_2 and NO_x emissions. It is important to note that the current energy system as sketched in Figure 1 closes the fossil fuel cycle through the atmosphere and that the costs of fossil fuel combustion are externalized.

The principal idea behind the novel, zero-emission energy system proposed here and illustrated in Figure 2 is to avoid closing the fossil fuel cycle through the atmosphere. Unlike the vertically structured sectors of the existing energy system, the new system would be composed of a horizontally integrated system consisting of four distinct stages. Previously independent sectors such as the coal, oil, or gas industries are merged into a interconnected system.

The principal idea of this proposed integrated energy system (IES) is to decompose and purify all fossil fuels at the front end of the energy chain. The net result of this process would be a variety of clean chemical products. These products would then be stoichiometrically supplemented for conversion or synthesis into energy forms needed to meet final energy requirements.

Specifically, the system starts with a number of energy and nonenergy inputs. According to Figure 2, the system's inputs include coal (solids or oil residues), natural gas, fissionable material, water, and air. Outputs of this decomposition and cleansing stage are the clean intermediate energy carriers carbon monoxide (CO), hydrogen (H_2), and oxygen (O_2). The technological processes deployed at this stage would comprise partial oxidation of solids using pure oxygen from air separation plants to form synthesis gas (a mixture of CO and H_2). High-temperature combustion would help avoid the generation of complex hydrocarbon compounds, sulfur emissions, etc. The molten iron bath process is one suitable technology for this purpose.

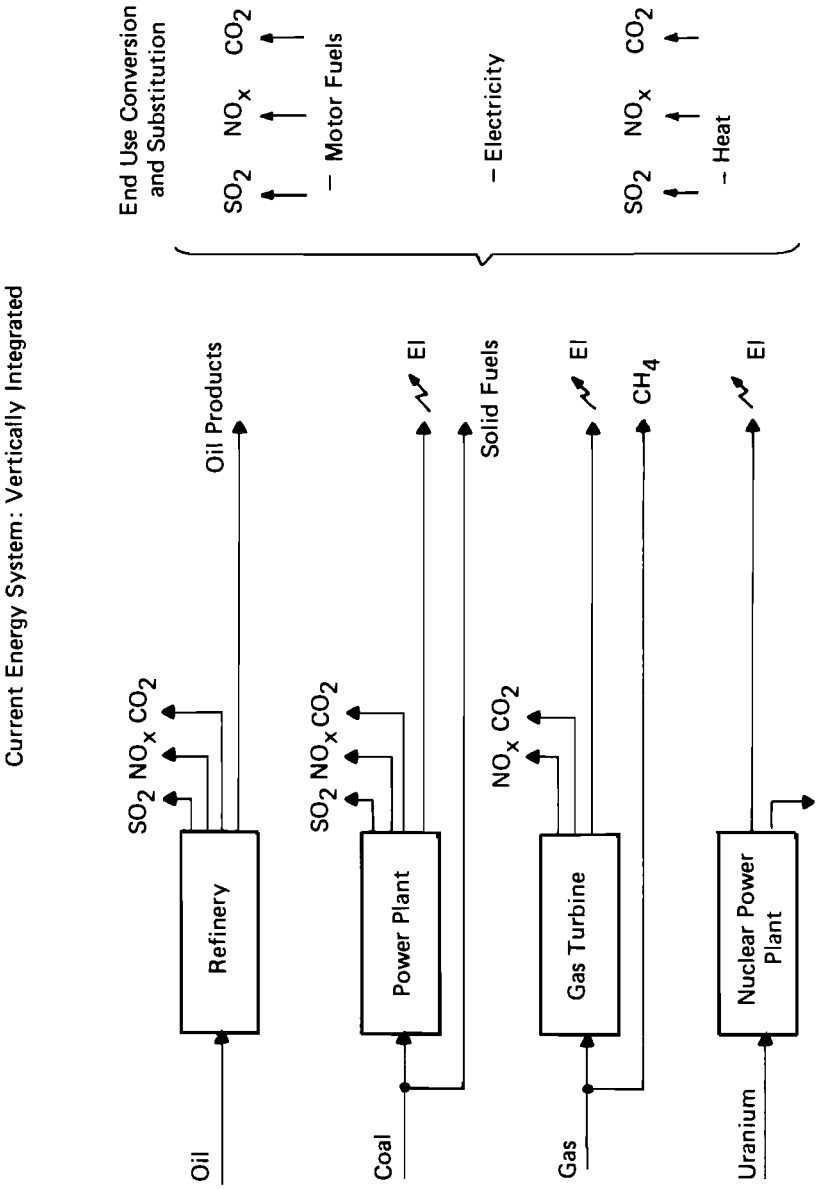


Figure 1 Current energy system, vertically integrated
Source: Häfele et al. (1986)

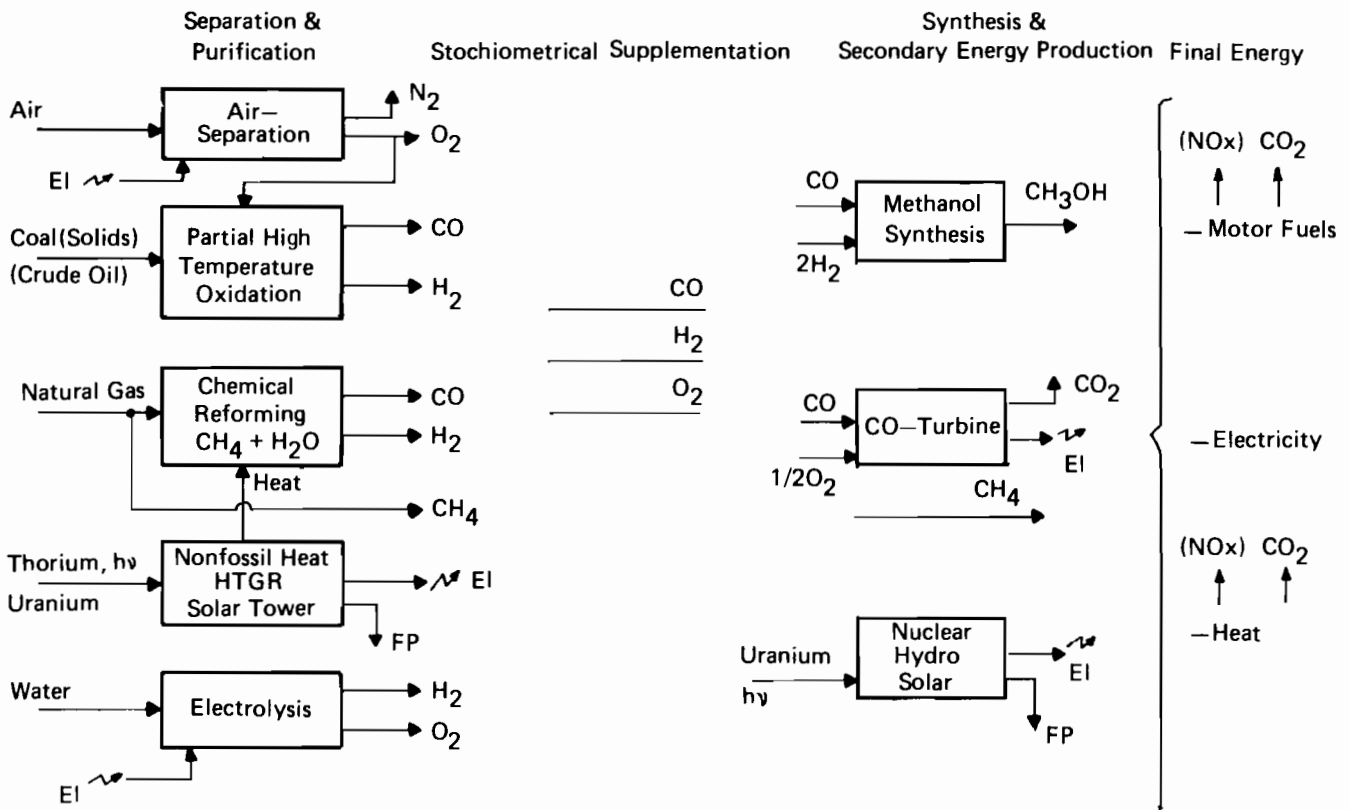


Figure 2 Novel energy systems, horizontally integrated
Source: Häfele et al. (1986)

Rather than deploy the inefficient operation of the combusting part of the CO to split water (the net effect of all gasification, liquefaction, and cracking schemes), the strategic choice in the IES would be different. Natural gas or methane would supply the stoichiometrically missing amounts of hydrogen for the synthesis of liquid fuels by means of methane reforming. Another possibility for the provision of additional hydrogen (and oxygen) foreseen in this configuration is off-peak electrolysis.

Chemical reforming of natural gas with steam yields synthesis gas plus one additional molecule of hydrogen. Steam reforming is an established technology that operates with clean substances and is thermodynamically attractive: it provides an inlet for high-temperature heat in the range of 700 to 900°C, which can split the water molecule at the expense of decomposing methane. Conceivably, rather than apply natural gas process heat for the shift reaction (which is both resource consumptive and environmentally polluting) high-temperature reactors or solar tower plants can be the source of process heat. In this way, methane reforming can serve as an interface to the non-fossil part of integrated energy systems.

The second stage of such a system consists of the stoichiometric complementation/supplementation of the decomposed and purified intermediate energy products. This step controls the mass flows of CO, H_2, and O_2 and directs the flows to the different conversion and synthesis processes (the third stage). The various forms and quantities of final energy demand determine the mode and operation of the technologies in stage three, the outputs of which are electricity, synthetic liquid fuel (here considered to be methanol), and natural gas. Note that stage three also includes "conventional" electricity production by hydropower, nuclear, and solar sources which are traditionally not associated with CO_2, SO_2, and NO_x emissions. Similarly, natural gas is piped through the system to the sites of end-use conversion in a conventional fashion.

A newly introduced technology to this configuration is the CO turbine. In combining with pure oxygen, this turbine is ideally suited for peak-load electricity production. The isothermal expansion of both CO and O_2 yields high efficiencies of the order of 60% as well as a significant reduction in generated CO_2. This scheme could operate without emitting any stack gases. The combustion product, CO_2, would be contained and used, say, for enhanced recovery in crude oil production, for other chemical needs, or in liquid form as a convenient solvent to absorb other waste products for storage in depleted gas fields, etc.

The configuration of the novel energy system, as shown in Figure 2, eliminates the need for intermediate electricity production by utilizing off peak nuclear or other non-fossil electricity for specially designed low capital cost electrolysis to contribute sufficient hydrogen and oxygen to the overall system. The present competition between coal and nuclear power in electricity generation subsystems would be gradually resolved and coal would become competitive where it matters: as a substitute for oil. Further, one should note that all major conversion facilities produce certain amounts of waste heat which can easily be supplied to district heat systems and hence further improve the overall system performance.

The conversion of synthesis gas to a liquid fuel is another strategic technology of the third stage. Methanol appears to be a prime candidate, at least as an intermediate form of energy for long-distance transportation and as a substance for blending classical hydrocarbons.

The fourth stage of the novel energy system is the end use conversion part of the energy chain. Given the configuration of Figure 2, the conversion of the various forms of final energy into energy services is accompanied mainly by carbon dioxide emissions (and by at least two-thirds of nitrous oxides emissions than in the conventional energy system).

The full advantage of this novel energy system depends on the establishment of a few interconnections, namely H_2, O_2, CO and CH_3OH-pipelines and storage capacity. These grids would introduce the necessary flexibility and a certain robustness to the global

energy system and mobilize erstwhile inaccessible energy opportunities. The most valuable characteristic of the novel energy system is the decoupling of the production of liquid fuels from any particular fossil fuel resource. By varying the input ratio of methane, nuclear and other non-fossil energy, to the input of the carbon rich components, any changes in the availability or market price of the latter can be compensated for. At the same time one obtains an environmentally desirable energy system characterized by an overall reduction in pollution emissions unprecedented to date. This occurs without giving up any opportunities, be it with respect to energy recources, technologies, or economics.

3. NOVEL ENERGY SYSTEMS AND METROPOLITAN REGIONS

The foregoing paragraphs presented one of many conceivable configurations of a novel energy system. The implementation of synergistic systems, however, is subject to a number of conditions, many of which are non-existent today. First of all there is the question of technological feasibility. Most of the components shown in Figure 2 are technically feasible and ready for immediate installation. Some technologies will be available in the foreseeable future. For the time being these could easily be substituted by existing processes, though at a reduced economic viability. The real obstacles, however, concern the institutional structure of today's energy industries and current energy policies. The novel system implies a total restructuring of the energy industries towards one essentially integrated industrial sector. The coal, oil, and gas industries are vertically structured (see Figure 1) and are competing against each other for market shares. They are far from joining forces for the benefit of an environmentally clean energy supply system. Instead, these industries are more likely to improve their existing vertically structured business once environmental regulations impose stricter emission standards (e.g. by installing abatement measures, etc.). Further, there is the complicated question of financing, benefit distribution, systems control and so forth. To that extent the proposed system appears to lack economic attractiveness, particularly in the short run and at energy price patterns prevailing in the mid 1980s.

The economics of a zero-emissions energy system change substantially if one internalizes all costs associated with fossil fuel combustion. Specifically, the external damage caused by burning a ton of coal or oil is estimated to be over $100 (Working Consulting Group, 1986). Since these costs are not part of standard utility accounting systems, it will take political action to include social and environmental damage costs into economic comparisons of future energy alternatives. Then the economics of an integrated energy system appear to be quite favorable compared with conventional alternatives.

Although the stylized integrated energy system shown in Figure 2 is a long-term concept, some of its basic principles have been implemented already. Electricity, gas or district heating networks are points in case. Because of environmental and efficiency considerations, the concept of central co-generation conversion plants has been favored in many densely populated regions, particularly in the northern hemisphere. To that extent, the integration of some energy systems has already begun. As already mentioned, the configuration in Figure 2 is only one of many conceivable integrated systems. Therefore, there are many alternative options and strategies for their implementation. In the following, one example of an energy system's integration will be presented. In the early 1980s this system was examined by the planning agencies, utilities and regulatory bodies of the larger Stockholm metropolitan region.

Apart from the institutional and policy considerations, any implementation of an integrated energy system within the geographical borders of a metropolitan region must take into close account the existing urban infrastructure; must recognize regional and national energy policies and must account for socio-economic constraints.

The study *Long-term Energy Supply Strategies for Stockholm County* (Temaplan, 1982) focused on the transition from an urban energy system, primarily based on oil, towards a non-oil dependent system. Another objective of the study concerned the reduction of the high energy import dependence of the Stockholm energy system and the associated risk of vulnerability caused by unexpected events in international energy markets. Apart from international energy markets, Stockholm County also depends on the Swedish national energy supply system. In particular, Stockholm County purchases electricity from indigenous Swedish hydroelectric and other power generation facilities as well as oil products from Swedish refineries.

It is the Swedish electricity generation sector which led to controversial deliberations in the late 1970s. The forecasts of national electricity demand in the 1970s were highly overestimated and have led to an excess electricity production capacity in Sweden. This abundance of production capacity is expected to prevail until the late 1990s. Thereafter, a serious squeeze on electricity supply is expected: Firstly, by that time the slowly growing electricity demand will surpass existing production capacities. Secondly, according to the last referendum on the peaceful utilization of nuclear power, the construction of additional nuclear power plants is prohibited (except two plants already under construction). Thirdly, the service time of existing nuclear power plants is limited to 25 years. This means that nuclear power stations which started production in the early 1970s will have to go off line during the late 1990s. Furthermore, the traditional source of electricity generation in Sweden, hydroelectric power, has almost reached its ecologically desirable limit.

Thus, the Swedish electricity sector faces a twofold problem with direct repercussions on the overall energy supply system of Stockholm County. The first problem is of a rather short term nature (the next 10 to 15 years or so) and concerns the question of the most economical use of the current excess electricity production capacity. The second question relates to the general long-term electricity supply structure.

The dilemma between the short versus the long term problem can be stated as follows: The excess electricity supply capacity suggests an intensive promotion of electricity usage so as to enhance sales and to minimize the surplus capacity gap. One promising market capable of absorbing additional electricity is the heating sector. This market is still oil dominated in low energy consumption areas, and offers good opportunities for electricity penetration as part of the national oil diversification strategy. However, given the constraints mentioned for the post 1990 period, the promotion of electricity sales through appropriate pricing policies will aggravate the envisaged long term problems of electricity supply.

One alternative to expanded electricity sales which was strongly promoted by the owners of underutilized power stations concerned the retro-fitting of power plants for district heat supply. This alternative has one serious draw-back. By the turn of the century the so converted capacities would not be available to meet electricity demand. The plan to switch these power stations back to electricity production (by the end of the 1990s) then raises the immediate question of future district heat supply alternatives.

The temporary utilization of the Forsmark 3 nuclear power station for district heat generation has a direct impact on the energy future of Stockholm. Presently the Forsmark 3 power plant is under construction and its completion is scheduled for the year 1984 (situation in 1981, when this study was performed). The current electricity demand outlook suggests that there will be no need for this power station before the mid-1990s. One alternative to recoup some of the capital investments before the turn of the century involves the retro-fitting of the Forsmark 3 power station for co-generation of electricity (300 MW(el)) to be supplied to the national grid and district heat (2000 MW(th)) for Stockholm county. This split between electricity and district heat reflects the ownership structure of the Forsmark 3 power station.

This alternative implies the construction of a heat transmission system (pipe lines) over a distance of 120 km from the Forsmark 3 location to the Stockholm County area, as well

as the necessary infrastructure, i.e. pump stations, heat exchangers and a 480 MW(th) oil-fired heating block back-up capacity. The retro-fitting would be completed by 1988; the switch back to electricity production is envisaged for the year 1998. This leaves a service time of roughly 10 years during which the capital investments for the retro-fitting and the peripheral equipment should have paid off.

After 1998, two 500 MW(el)/1000 MW(th) coal-fired co-generation blocks located in the vicinity of the Forsmark site are planned to balance the drop in district heat production capacity caused by reverting the 2000 MW(th) of Forsmark 3 to electricity generation. The heat transportation system, therefore, would be utilized beyond the year 1998.

In light of the serious energy import dependence of Sweden and Stockholm County in particular, it is desirable that the future energy system should be marked by intensified domestic energy production or at least by a higher import diversification. The Forsmark scenario implies a strong shift from oil to coal imports, especially for the period after the temporary nuclear era in Sweden. The target of import diversification has been achieved to a certain degree, since the potential coal exporting countries are more evenly distributed globally and belong to politically stable nations.

However, there are alternatives to the Forsmark scenario which, besides economic advantages, offer an even higher degree of energy import diversification. Such alternatives cannot emerge by analyzing the heat production sector in isolation. Only if the entire energy system is included in the analysis, will one be able to distil solutions which optimize metropolitan energy systems subject to the constraints of urban infrastructures.

For example, the fuel needs of the transportation sector in Stockholm County amount to roughly 20% of the total oil product consumption. An energy supply strategy, which in principle provides district heat equivalent to the Forsmark scenario, but at the same time substitutes for oil and oil product imports within the transportation sector, must be considered superior to the Forsmark scenario. The Nynäshamn Energy Complex represents such a promising alternative. Essentially, the main output of this complex would be methanol produced from coal. In the short run, methanol is a suitable synthetic fuel admixture to eke out traditional gasoline (and thereby oil) imports or to substitute fully for gasoline in the longer run. The methanol production process requires significant cooling requirements. The integration of this "waste heat" into the Stockholm County district heat system would put an otherwise vented by-product to practical use. To that extent, the Stockholm district heat market serves as the "cooling tower" for the methanol plant. The potential heat deliveries from this complex, however, would not suffice to meet total demand. Here, another technical feature of the methanol process foreseen at Nynäshamn offers an interesting solution. Typically, the return water of district heat systems has a temperature of some 60°C. The feed water temperature for the methanol process should be in the order of 20°C. The implementation of electricity driven heat pumps could utilize the 40°C temperature difference by extracting the heat and cooling the return water down to the desirable temperature of 20°C.

Up to the turn of the century, the electricity required to operate these heat pumps could easily be supplied by the Forsmark 3 power station. In the long run, adequate co-generation power plants based on coal (or natural gas) would have to replace the nuclear electricity generation.

The planned technical configuration of the methanol production plant at Nynäshamn also increases the flexibility of the Stockholm energy supply system. The proposed methanol production process consists of two major stages. In the first stage, a carbon source (coal) is converted into a fuel gas (gasified). In the second step, this fuel gas is further processed and synthesized to methanol. Therefore, it is technically feasible to divert fuel gas after the gasification step and to use this synthetic fuel gas in a very conventional way: e.g. for heating or cooking purposes in residential areas, as a boiler fuel for industries or as a fuel in co-generation plants (e.g. peak-shaving).

4. THE SPATIAL AND TEMPORAL FRAME

The spatial frame of this study was determined *a priori* - the Stockholm County. However, this is only the case as long as the future evolution of energy demand is concerned. Most energy supply facilities and primary energy resources, such as the Forsmark 3 plant or hydroelectricity generating plants, are located outside the Stockholm County area. The model, therefore, has to account for all energy production and conversion facilities which affect Stockholm County irrespective of their geographical location. This also requires the inclusion of an adequate model representation of the energy transport system from energy conversion plants outside the County to the Stockholm area, and the subsequent distribution inside the Stockholm area. Hence, the spatial frame of the Stockholm energy study (model) includes the Stockholm County area, the power plants which deliver secondary forms of energy to Stockholm and the connecting transport grids.

The temporal frame covers 40 years, i.e., the period between 1980 and 2020. The analysis of future energy supply options for Stockholm County requires an extended study horizon. The service time of energy conversion equipment or energy distribution equipment ranges from 10 years for residential furnaces to 50 years and more (hydroelectric power plants etc.). This is to say that quantitative considerations for energy planning purposes should be based on time horizons which are sufficiently long enough to include the effects of their implementation. Only then will energy alternatives which are flexible enough to respond to changes in the world energy situation emerge and which are therefore characterized by a certain robustness against unexpected events.

5. THE BASIC STRUCTURE OF THE ENERGY SUPPLY MODEL MESSAGE II

MESSAGE II is a dynamic linear programming model which reflects the essential stages of energy chains from primary energy extraction to end-use conversion (Messner, 1984; Strubegger, 1984). Given the absolute level and the types of useful energy demand, the model calculates the cost-optimal energy supply strategy for a predetermined geographical area (metropolitan areas, regions or national economies, etc.). This cost-optimal strategy is subject to a number of constraints. These constraints limit the number of conceivable trajectories towards a future energy supply system and force the model's solutions to remain within the range of plausibility. The first set of constraints serves the correct representation of the current status/infrastructure of the energy system under scrutiny. Other constraints restrict or define the dynamics of change, the time profile for the availability of new technologies or politically determined energy import ceilings.

In the Stockholm study, special emphasis was given to a detailed representation of end-use conversion technologies, particularly energy saving technologies. Altogether 38 different heating technologies were implemented to analyze the future options for meeting the useful energy demand for heating purposes. Furthermore, the model adjusts useful energy demand levels according to endogenously calculated final energy supply costs. Technology and/or interfuel substitution are thus a direct consequence of relative fuel price changes at the burnertip. To that extent, energy price elasticities are explicitly implemented in MESSAGE II.

MESSAGE II also hosts the possibility of modeling mixed integer problems, which is of particular importance when analyzing distinct energy production or transport facilities. For example, the Forsmark-Stockholm heat transport pipeline has only one technoeconomically feasible configuration. The decision therefore can be either go or no-go, but not a linear combination such as only using two-thirds of the diameter of that transport pipeline.

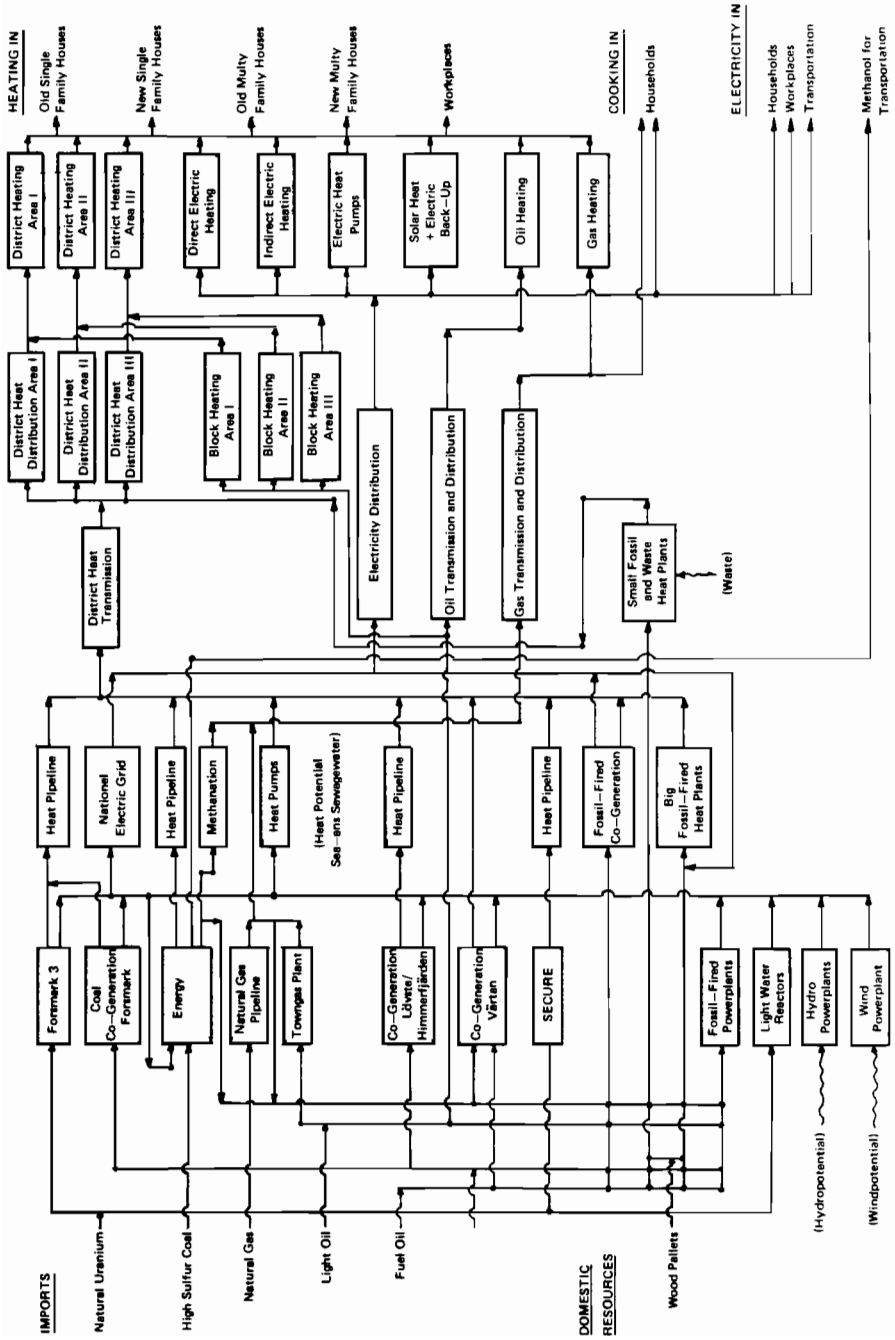


Figure 3 A schematic representation of the energy system of Stockholm County

Figure 3 shows a simplified, schematic representation of the current energy infrastructure of Stockholm County including all the future options which have entered the recent energy debate in Stockholm. The Värtan co-generation plant, the boxes labelled fossil power or small-scale heat plants are examples of existing technologies, while Forsmark 3 and the Nynäshamn energy complex represent future options. Various transmission and distribution systems for the different types of secondary energy complete the representation of Stockholm County's energy system.

6. SOME PERTINENT RESULTS

The Stockholm energy study focuses on the technoeconomic evaluation of two principal alternatives viz. the Forsmark district heat supply and the Forsmark electricity/Nynäshamn complex scenarios. Some 25 scenario variations were analyzed to account for the uncertainties associated with future fuel import prices, methanol revenues, district heat tariffs, the degree of the integration of Stockholm's electricity sector into the national grid, etc. Despite often large variations regarding basic scenario assumptions, the numerical results of the model applications show some remarkably consistent, and therefore robust, results. For example, the heating structure in single and multi family homes turns out to be almost independent of the primary energy supply configuration. Figures 4 and 5 show the uniform, i.e. scenario independent, phase-out of oil consuming heating systems. During the initial time period, the implementation of energy savings measures to the pre-1975 constructed building stock appears to be the economically most attractive response to the rising (or constant) oil market prices. Clearly, in the short run the end use part of the energy systems is more flexible in responding to market disruptions than the energy production system.

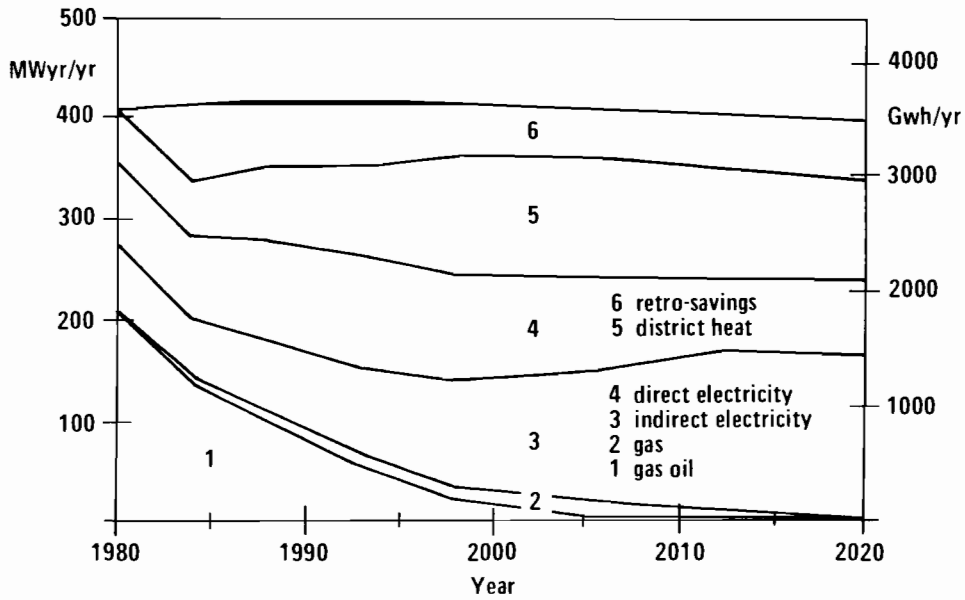


Figure 4 Stockholm County: Single family houses, structure of space heating

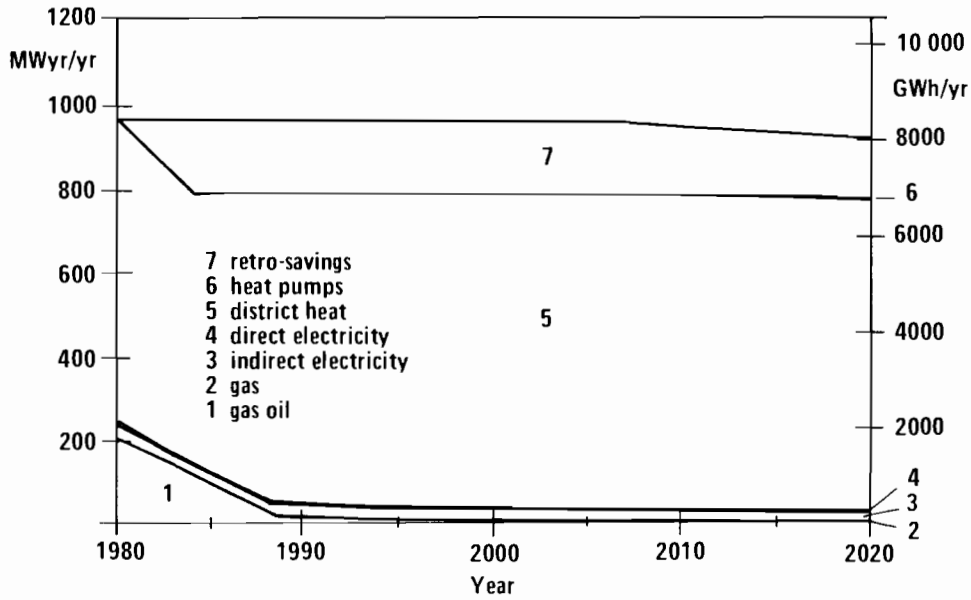


Figure 5 Stockholm County: Multi family houses, structure of space heating

The final energy demand evolution for single-family houses reveals interesting aspects. The significant plunge of final energy requirements will be partly reversed by the end of the 1980s. This points to the fact that two components - different in nature and with diverse implications in the long run - cause the reduction in final energy consumption during the initial period. The first component concerns the immediate reaction of households to fuel cost increases by cutting energy consumption - e.g. by turning down thermostats., etc. - which is essentially price-induced energy conservation. The second component concerns the substitution of capital for energy, i.e. better insulation of homes, replacement of inefficient furnaces or interfuel-substitution. The latter component bears a long-term effect on overall energy consumption. Landlords are not likely to reverse their investment decision because of unexpected lower fuel price developments. The price-induced reduction in energy use, however, is reversible; and that is exactly what happens in Figure 4. According to the scenario specification, major electricity and district heat production capacities will be taken into operation during the period 1984 to 1988. Due to the surplus capacities in the national electricity sector (and the fact that these additional capacities involve nuclear power plants primarily), the shadow price for base load electricity drops sharply as indicated in Table 1. At the same time, large supplies of district heat become available (either from Forsmark or the Nynäshamn energy complex), at costs lower than oil fired heating systems. Thus, the incentives for price induced conservation vanishes and demand rises to previously observed levels (minus the savings achieved through the capital for energy substitution).

In all building types, oil use for heating purposes is reduced drastically and by the year 2010 oil disappears from Stockholm's heat supply menu. In low energy consumption density areas, single family homes substitute electricity based heating systems for oil-fired furnaces. In the higher density areas, district heat competes successfully against oil and electricity. Multi-family houses, traditionally the focus of district heat supply systems, continue to displace oil and switch to a centralized heat supply. In the commercial/small

Table 1 Total district heat production (MWyr/yr), mixed integer scenario.

Technology of Fuel Used	1980-1983	1984-1987	1988-1992	1993-1997	1998-2004	2005-2011	2012-2019	2020
Waste	40.00	70.00	56.55	56.55	48.47	0.00	85.60	85.60
Oil	596.85	557.60	178.32	143.51	25.04	23.62	17.54	18.76
Coal	0.00	296.99	173.09	79.70	79.70	9.11	0.00	0.00
Wood	0.00	0.00	0.00	0.00	0.00	0.00	30.73	48.00
Electric boilers	20.35	3.84	0.00	0.03	0.00	0.00	0.00	0.00
Total heat plants	657.20	928.43	407.95	279.79	153.21	32.73	133.87	152.36
Oil	186.30	65.91	13.77	13.77	2.18	0.00	0.00	0.00
Coal	0.00	88.00	264.51	555.33	1134.48	1337.72	1287.95	1315.02
Gas	0.00	0.00	37.30	43.41	16.37	15.22	4.20	3.81
Värtan	59.01	81.71	59.17	27.58	15.37	4.75	4.75	4.75
Total cogeneration	245.31	235.62	374.75	640.08	1168.40	1357.69	1296.90	1323.58
Energy complex	0.00	0.00	487.47	472.01	351.34	337.30	291.80	265.72
Heat pumps	0.00	0.00	205.67	204.44	44.23	68.81	99.63	124.15
Total special	0.00	0.00	693.14	676.45	395.58	406.11	391.43	389.88
Total	902.50	1164.05	1475.84	1596.32	1717.19	1796.53	1822.19	1865.82

industry sectors, the share of oil is also replaced primarily by district heat. Some contributions from electric and solar technologies complement the heat supply in these sectors.

By the year 2020, district heat accounts for more than 80% (25% in 1980) of the Stockholm heat market and for 50% of total final energy consumed in Stockholm County (excluding the transport sector). The remaining 50% of final energy consumption is electricity. This development means a considerable reliance on grid dependent energy carriers and - as will be shown - points to a shift toward an integrated energy system.

All long-term energy supply options for Stockholm County analyzed in this study are also marked by a supply side commonality. By the end of the study horizon, i.e. the year 2020, district heat production and electricity generation is almost identical in all scenarios. This is certainly not an entirely unexpected result, given constraints such as the institutionally enforced restrictions on the use of nuclear power and the poor energy resource situation of Sweden and Stockholm County in particular. In addition, the objective of reducing the serious oil import dependence quasi predetermines the principal feasible solution space which meets these objectives and constraints.

Table 1 depicts the district heat production structure for the period 1980 to 2020.[2] The data of Table 1 illustrate the dynamic transition from oil heating plants to coal-fired co-generation plants common to all scenarios calculated. In 1980, almost all district heat production is based on oil, one-third of which is burnt in co-generation plants. The sharp increase in district heat demand during the mid-1980s is met by coal fired heating plants. Due to the existing surplus capacity of the national electricity sector, only a few co-generation capacities are installed. By and large oil defends the 1980 market share

[2]The scenario underlying Table 1 is the energy complex scenario which turned out to be the cost-optimal solution. Forsmark 3 is used for electricity generation exclusively.

throughout the 1980s. After 1990, oil looses economic attractiveness in both centralized and decentralized heat markets.

Toward the end of this decade, the energy complex starts operating and supplies one-third of Stockholm's district heat needs. The rapidly expanding demand for district heat causes a squeeze in the Stockholm energy system, and co-generation based on coal is being introduced at an increasing rate. The electricity output of the co-generation plants is partly used for the operation of large heat pumps. Depending on the load variation (seasonal and daily), and the capacity utilization structure of the national electricity production sector, other technologies contribute varying quantities to Stockholm's heat supply. By the end of the study horizon, the heat supply structure is dominated by coal-fired co-generation plants supplying more than 70% of Stockholm's heat consumption. The remaining 30% are covered by the energy complex, heat pumps and heating plants, the latter of which primarily supply peak demand. As already mentioned, after the year 2010 the structure of district heat production (and electricity generation) is almost identical in all scenarios.

The long-term structure of Stockholm County's energy system is quite similar for all scenarios and shows robustness with regard to considerably diverse short-to-medium term trajectories. Therefore, it appears reasonable to use the overall cost-effectiveness of the energy system as decision criteria for selecting the most appropriate trajectory. Obviously, criteria other than pure cost considerations - in particular, environmental aspects - supplement this cost-optimal approach.

Among the numerous proposals concerning the utilization of the Forsmark 3 nuclear power station, the criterion of cost-effectiveness rejects the district heat alternative. Instead, the energy complex proves to be competitive even under very conservative assumptions regarding sales revenues from methanol or future oil prices. The objective of energy import diversification is met in all scenarios analyzed. Again, the largest energy import diversification occurs in those cases where the energy complex participates in the cost-optimal solution. In summary, the Stockholm County energy analyses have shown that the criteria of cost-effectiveness and long-term robustness of the energy system in regard to unexpected events strongly favors the construction of the energy complex at Nynäshamn.

REFERENCES

Häfele et al., 1981, *Energy in a Finite World: Vol. 2 A Global Systems Analysis,* Report by the Energy Systems Program Group of IIASA, Ballinger, Cambridge, Mass.

Häfele, W., H. Barnert and W. Sassin, 1982, Künftige fossile Brennstoffe: Ihre Nutzung und Einbettung in moderne Energiesysteme, Nuclear Research Center Jülich, FRG.

Häfele, W., H. Barnert, S. Messner, M. Strubegger and J. Anderer, 1986, "The Concept of Novel Horizontally Integrated Energy Systems: the Case of Zero Emissions", in W.C. Clark and R.E. Munn (eds.), *Sustainable Development of the Biosphere,* Cambridge University Press, Cambridge.

Meadows, D., D. Meadows, E. Zahn and P. Milling, 1972, *The Limits to Growth.* A Report for the Club of Rome's Project on the Predicament of Mankind, Deutsche Verlags-Anstalt, Stuttgart.

Messner, S., 1984, "Users Guide for the Matrix Generator of MESSAGE II, Part I: Model Description and Implementation Guide; Part 2: Appendices", Working Papers WP-84-71a and WP-84-71b, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Rogner, H.H., 1982, "Substitution of Coal and Gas for Oil - some Global Considerations". Paper presented at the First US-China Conference on Energy, Environment, and Resources, Beijing, Nov. 7-12, 1982, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Sassin, W., 1982, "Fossil Energy and its Alternatives - a Problem Beyond Costs and Prices". Paper presented at the International Economic Association Conference on Economics of Alternative Soures of Energy, Tokyo, Sept. 7 - Oct. 1, 1982, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Strubegger, M., 1984, "User's Guide for the Post-processor of MESSAGE II", Working Paper WP-84-72, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Temaplan, 1982, Long-Term Energy Supply Strategies for Stockholm County, Stockholm, TEMAPLAN GmbH, Böblingen, FRG.

Working Consulting Group of the President of the Soviet Academy of Sciences on Long-Term Energy Forecasting, M.A. Styrikovich, Research Leader, 1987, International Natural Gas Market, Working Paper WP-87-102, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Advances in Spatial Theory and Dynamics
A.E. Andersson et al. (editors)
Elsevier Science Publishers B.V. (North-Holland)

CHAPTER 19

Taste Changes and Conservation Laws in the Housing Market

K. Kobayashi, W.B. Zhang and K. Yoshikawa

1. INTRODUCTION

Recently, the issue of social and economic structural changes in urban areas has received considerable attention from researchers in urban economics, regional science, and geography. Structural changes have generally been considered to result from changes in technology and the tastes of human beings. Some studies try to explain structural changes according to variant tastes and other variant factors such as progress in information and technology. Bifurcation theory has been employed to capture structural changes in socio-economic systems (see, for example, Wilson, 1981) as well as many other fields in regional science (Wilson, 1981; Barentsen and Nijkamp, 1986). On the other hand, the Brussels school employs a different approach to urban structural changes through the use of dissipative systems theory (e.g., Allen, 1985).

In all of these works, structural changes are described by models in which parameter shifts are exogenously given. The parameters in the models are assumed, explicitly or implicitly, to be related to technology, tastes of human beings, and other factors which have impacts upon variables in the models, but are exogenously determined. However, parameters in the models are usually measured by observing the behavior of human beings, which results in difficulties in predicting changes of parameters in the future. Yet, in some situations, changes in technology and tastes can directly influence "effective" values of variables in the models. In these situations, the methods mentioned above can not effectively capture the characteristics of the changes in progress. In order to measure effects of changes in technology and tastes directly, a different way of coping with structural changes needs to be developed.

In this paper, a systematic way of dealing with the taste changes of human beings shall be discussed. Instead of analysing how changes in parameters affect the behavior of models, our approach shall keep parameters constant, even under the influence of changes in the tastes of human beings. The effects of exogenous changes in tastes upon the variables in the model are measured directly by transformations satisfying Lie group properties. Our approach offers advantages in the sense that taste changes can be explicitly included in the model. By keeping structures of the model consistent, it is possible to take account of taste changes in analysing the behavior of human beings in our approach. We can also determine whether a model is invariant under impacts of taste changes, i.e., whether a model is "permissible" in the presence of taste changes. Although many forms of functions in existing urban models have been employed, there are different "permissible" patterns of taste changes for given forms of functions. In

practical terms, we need models such as these which are invariant under wider patterns of taste changes, since if a model is not invariant under impacts of taste changes, behavior derived from the model is very different even for infinitesimal changes in tastes.

In this study we use infinitesimal transformations in Lie group theory to express taste changes of households. We shall also explicitly define the invariance of dynamic models under Lie groups. The models which are invariant under taste changes shall be investigated. It will be noted that although Lie group theory has been applied to theoretical economics by Sato (1981), to our knowledge there have been few applications of Lie group theory to regional economics and regional science. In this paper, we try to investigate the ability of Lie group theory to cope with taste changes of households through the analysis of the behavior of a developer.

In Section 2 we define a basic dynamic model for the behavior of developers under the perfect foresight hypothesis. Section 3 is devoted to economic interpretations of Euler's equations for the basic model. Taste changes of the households and infinitesimal transformations to describe the taste changes are explicitly defined in Section 4. In Section 5, we provide definitions of dynamic and divergent invariance for the basic model. The necessary conditions for the system to be invariant are derived in Section 6. By use of the Noether theorem we prove existence of conservation laws in the housing market in Section 7. Section 8 discusses structures of the housing market under taste changes of the households. In section 9 we present some concluding remarks about this study.

2. THE DYNAMIC MODEL

This paper is particularly concerned with behavior of a developer in a housing market undergoing urban change. The developer is assumed to be rational in the sense that he supplies houses to maximize the discounted sum of the utility (profit) stream in a perfectly competitive housing market. Although many dynamic models exist for developers in the housing market, Fujita (1983) has classified these dynamic models into four groups according to hypotheses about the behavior of the developers included in the models: (1) static expectation; (2) perfect foresight; (3) rational expectation; and (4) adaptable expectation. It can be said that most of the research efforts related to dynamic models have adopted the perfect foresight hypothesis (PF-hypothesis). Our model is an extension of the dynamic model based upon the PF-hypothesis presented by Diamond et al. (1982), which explicitly accounts for the developer's investment in housing.

The total profit of the developer at any time t is

$$F(t) = R(A(t),Q(t)) - C(A(t),\dot{Q}(t)), \tag{1}$$

where

$F(t)$ = the total profit at time t,
$A(t)$ = the level of amenities at time t, a vector variable,
$Q(t)$ = service values of houses at time t, a vector variable,
$\dot{Q}(t)$ = investment in service values of houses,
$R(A(t),Q(t))$ = revenue function at time t,
$C(A(t),\dot{Q}(t))$ = cost function at time t.

For simplicity we will not express time t explicitly in these variables in the following.

We suppose that the levels of amenities are controlled by the government and are not affected by the behavior of the developer. Amenities affect the behavior of the developer

because the revenue and cost of supplying houses depends upon the amenities. It is assumed that the developer decides the level of amenities by choosing the location of the houses at the initial stage. Once the developer chooses the housing site, the changes in amenities occur exogenously. According to the PF-hypothesis, the developer maximizes the discounted sum of utility stream for profit,

$$\text{Max} \int_0^{\infty} U(R(A,Q) - C(A,\dot{Q}))\exp(-kt)dt\,, \tag{2}$$

where

U = U(F(t)) = utility function,
k = discounting rate.

3. EULER'S CONDITIONS FOR THE BASIC MODEL

The Lagrangian for the basic model is defined as

$$L(A,Q,\dot{Q},t) = U(F(t))\exp(-kt)\,. \tag{3}$$

Euler's conditions for the optimal problem can be expressed as

$$\partial L/\partial A = d/dt(\partial L/\partial \dot{A})\,, \tag{4}$$

$$\partial L/\partial Q = d/dt(\partial L/\partial \dot{Q})\,. \tag{5}$$

With eq. (4)

$$U'\,\partial F/\partial A\,\exp(-kt) = 0\,, \tag{6}$$

where U' = dU/dF. From eq. (6), since U' is not equal to zero, the following equation holds

$$\partial R/\partial A = \partial C/\partial A\,, \tag{7}$$

which means that the marginal revenue of the amenities is equal to the marginal cost of the amenities. From eq. (5), we can obtain

$$U'\partial R/\partial Q = kU'\partial C/\partial \dot{Q} - d/dt(U'\partial C/\partial \dot{Q}). \tag{8}$$

Here, for simplicity we consider the situation that the developer is neutral to the profit, i.e., U(F(t)) = F(t). Then, for eq. (8),

$$\partial R/\partial Q = k\,\partial C/\partial \dot{Q} - d/dt(\partial C/\partial \dot{Q}). \tag{9}$$

Here, if we denote $\partial C/\partial \dot{Q}$ as p, eq. (9) can be rewritten as

$$\partial R/\partial Q = kp - dp/dt. \tag{10}$$

If there is no scale factor in the housing market, i.e., $dp/dt = 0$, from eq. (10) we have $\partial R/\partial Q = kp$, which means that the marginal revenue of the service values is equal to the marginal cost of the investment in service values discounted at rate k. In the case of a scale factor of the housing market, i.e., $dp/dt < 0$, the marginal revenue of the service values is greater than that without a scale factor. Here, let us denote $\partial R/\partial Q$ with r and take differentials of eq. (10) with respect to time t. We have

$$dr/dt = kdp/dt - d^2p/dt^2 . \tag{11}$$

In the case with scale factors, $d^2p/dt^2 > 0$, from which we get $dr/dt < 0$. Eq. (11) tells us that if there exists scale factors in the housing market, the marginal revenue of the service values will decrease as time passes.

4. TASTES OF HOUSEHOLDS AND INFINITESIMAL TRANSFORMATIONS

The tastes of households for locations and service values of houses are variant. The developer has to take these changeable factors into account in order to maximize the utility. It has been shown by Sato (1981) that Lie groups can express changes of tastes in "effective values" of variables in the model.

We have to use "units" to measure the values of these variables. The commensurate values of the variables can be compared with the units which are given a priori. The absolute values of the units are assumed to be invariant. However, taste changes may change the absolute values of the units. For example, the value of accessibility may change if the way in which the household values its time changes. We may say that tastes of households reflect how the households value variables related to subjective appreciation. It is reasonable to consider that changes in tastes can be described by changes in "effective values" of the variables.

We can measure the values of variables of a model in a local coordinate. As the meanings of "units" become different, the local coordinate is changed. These changes can be described by mapping between the local coordinates, utilizing the infinitesimal transformations in Lie group theory. Such transformations create maps that directly reflect changed measures of the effective variables valued at different points. The taste changes could be defined as transformations of effective values between different points in time. These ideas can be described by Lie groups expressing mapping between different local coordinates as shown in Figure 1.

Definition (Taste Changes of Households)

When exogenous taste changes are introduced, taste changes are expressed by transformation of the state space (A,Q,t). The transformation combines the factors before changes with the factors after changes due to the taste change parameter, which may be called taste changes of the households, i.e.,

$$TR_\varepsilon: \quad \bar{t} = Y(A,Q,t,\varepsilon) , \tag{12}$$

$$\bar{A} = X_1(A,Q,t\,\varepsilon) , \tag{13}$$

$$\bar{Q} = X_2(A,Q,t\,\varepsilon) , \tag{14}$$

where $\bar{t}$, $\bar{A}$, and $\bar{Q}$ are "subjective time" (Samuelson, 1976), "effective amenities" and "effective service values" respectively; (Y, X_1, X_2) is a vector of taste change functions; ε is a parameter to express the taste changes, and (Y, X_1, X_2) are assumed to have properties of continuous (Lie) groups with the parameter ε.

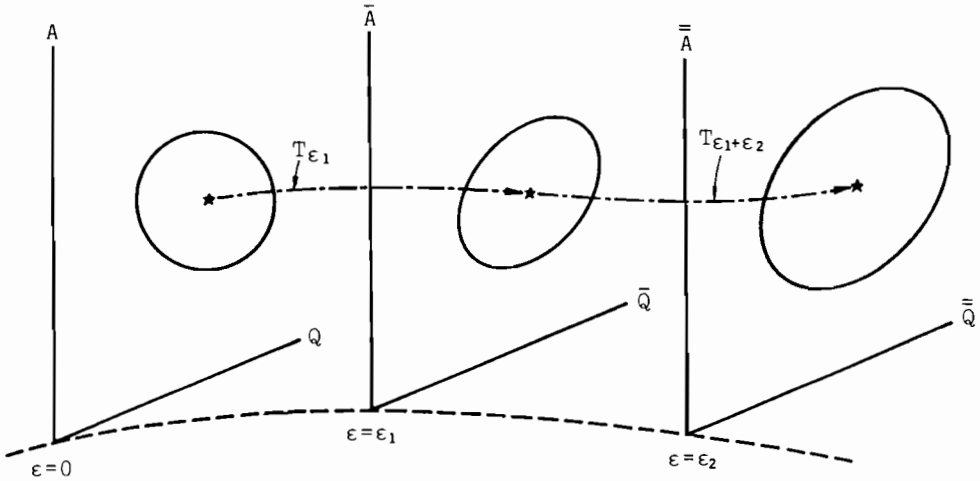


Figure 1 Taste Changes and the Local Coordinate

$\bar{t}$ can be considered as "subjective" time in the housing system which means a "real sense" of time. If we consider ε as an infinitesimal change in the normal time in our every-day economic life, $\tau = 1$. Because of taste changes of the households, causing value changes in the amenities and the houses, we may consider that the concept of time t in the housing system is different from that in our everyday economic life. Subjective time may be decided due to changes in tastes of the households in comparison with other economic sectors, if we consider the housing system as a sub-system in the national economy. When $\bar{t} = t+\varepsilon$ the concept of subjective time has no particular implication if we consider ε as an infinitesimal change in general time.

If we define the infinitesimal generators τ, ξ_1, and ξ_2, as

$$\tau = \partial Y(A,Q,t,\varepsilon)/\partial\varepsilon\,, \tag{15}$$

$$\xi_1 = \partial X_1(A,Q,t,\varepsilon)/\partial\varepsilon\,, \tag{16}$$

$$\xi_2 = \partial X_2(A,Q,t,\varepsilon)/\partial\varepsilon\,, \quad \text{at } \varepsilon = 0\,, \tag{17}$$

the expressions of taste changes in eqs. (12) to (14) are

$$\bar{t} = t + \tau\varepsilon + 0(\varepsilon^2)\,, \tag{18}$$

$$\bar{A} = A + \xi_1\varepsilon + 0(\varepsilon^2)\,, \tag{19}$$

$$\bar{Q} = Q + \xi_2\varepsilon + 0(\varepsilon^2) . \tag{20}$$

Here, if the amenities and service values are (A,Q) at the initial point ε=0, the real values of these variables can be considered as $(\bar{A},\bar{Q})$ at state ε as a result of taste changes of the households in the housing market. The inverse transformations of eqs. (12) to (14) are

$$TR_{-\varepsilon}: \quad t = Y(\bar{A},\bar{Q},\bar{t},-\varepsilon) , \tag{21}$$

$$A = X_1(\bar{A},\bar{Q},\bar{t},-\varepsilon) , \tag{22}$$

$$Q = X_2(\bar{A},\bar{Q},\bar{t},-\varepsilon) , \tag{23}$$

$$\text{or} \quad t = \bar{t} - \tau\varepsilon - 0(\varepsilon^2) , \tag{24}$$

$$A = \bar{A} - \xi_1\varepsilon - 0(\varepsilon^2) , \tag{25}$$

$$Q = \bar{Q} - \xi_2\varepsilon - 0(\varepsilon^2) , \tag{26}$$

by which the real values of the variables at state ε can be converted into those at the initial state ε_0. Inverse transformations can be utilized to transform the effective values of variables at any time into those at the base year. However, changes in taste are nothing more than changes of measures and there are no guarantees that the basic model will always remain dynamically consistent even under impacts of changes in effective values of the variables. Later we will explain that only limited families of the basic model are able to hold dynamic invariance under taste changes.

5. DEFINITIONS OF INVARIANCE FOR THE BASIC MODEL

In general, there are two concepts of invariance for a dynamic system under infinitesimal transformations, i.e., dynamic invariance and divergent invariance (Ikeda, 1975; Sato, 1982). Based upon these concepts, two concepts of invariance of the basic model can be defined as follows

$$L(\bar{A},\bar{Q},\dot{\bar{Q}},\bar{t}) - L(A,Q,\dot{Q},t) = 0(\varepsilon) , \tag{27}$$

or

$$L(\bar{A},\bar{Q},\dot{\bar{Q}},\bar{t}) - L(A,Q,\dot{Q},t) = \varepsilon\Omega(A,Q,t) + 0(\varepsilon) , \tag{28}$$

which mean dynamic invariance and divergent invariance under taste changes respectively.The dynamic and divergent invariance can be interpreted as necessary conditions for optimality of the problem under taste changes. In (28), Ω is an arbitrary function.

The Lagrangian, $L(A,Q,\dot{Q},t)$, denotes the discounted utility level at state ε. Let us represent the values of a house at initial state $\varepsilon = 0$ as (A,Q), and the values at state ε as $(\bar{A},\bar{Q})$. $(\bar{A},\bar{Q})$ are "effective values" after the tastes of households have changed. Since $(\bar{A},\bar{Q})$ and (A,Q) characterize the same house, when ε is infinitesimal, the discounted values at the state measured by $(\bar{A},\bar{Q})$ must be different from those at $\varepsilon = 0$ measured by (A,Q) to a infinitesimal amount $0(\varepsilon)$. The dynamic invariance denotes that even when taste changes occur, we can still use the same revenue and cost functions in the model to explain the behavior of a developer in the housing market. On the other hand, the

divergent invariance states that when taste changes arise, but the utility function retains the same form, the discounted utility level is shifted. In practice, not only tastes, but also other factors modify behavior of the developer. The divergent invariance reflects that when the housing system is affected by exogenous factors, the discounted utility levels can be modified. However, the form of the utility function need not be changed. In any case, if the basic model is dynamic or divergent invariant for certain classes of taste changes, the properties deduced from the basic model, e.g., optimal conditions, can remain invariant.

6. NECESSARY CONDITIONS FOR THE SYSTEM TO BE INVARIANT

Above we have defined dynamic and divergent invariance under impacts of taste changes for the basic model. Whether the system is invariant depends upon the forms of taste changes and rent and cost functions included in the model. That is, forms of the model on the basis of the PF-hypothesis are limited if the system is kept invariant under impacts of taste changes. In the following, we shall investigate necessary conditions for the housing system to be invariant. If the fundamental integral (2) is invariant under Lie groups, the corresponding Lagrangian must satisfy the following condition

$$\tau\partial L/\partial t + \xi_1 \partial L/\partial A + \xi_2 \partial L/\partial Q + \partial L/\partial \dot{Q}(d\xi_2/dt - \dot{Q}d\tau/dt) + Ld\,\tau/dt = d\,\Omega/dt\,, \quad (29)$$

which is called the fundamental invariant identity for the dynamic system (Sato, 1981). In the dynamic invariant system, the fundamental invariant identity can be written as

$$U(d\tau/dt - k\tau) + U'(\xi_1 \partial R/\partial A - \xi_1 \partial C/\partial A + \xi_2 \partial R/\partial Q - \partial C/\partial \dot{Q}d\,\xi_2/dt + \dot{Q}\partial C/\partial \dot{Q}\,d\tau/dt) = 0\,. \quad (30)$$

The conditions that eq. (30) is held for any U and U' except U = 0 and U' = 0 are given by

$$d\tau/dt - k\tau = 0\,, \quad (31)$$

$$\xi_1 \partial R/\partial A - \xi_1 \partial C/\partial A + \xi_2 \partial R/\partial Q - \partial C/\partial \dot{Q}d\,\xi_2/dt + \dot{Q}\partial C/\partial \dot{Q}\,d\tau/dt = 0\,. \quad (32)$$

Since the generator τ is a function of A, Q and t, the general solution of eq. (31) can be written as

$$\tau = a\exp(kt) + \psi(A)\,, \quad (33)$$

where $\psi(A)$ is an arbitrary function of A. Similarly, from eq. (32) we obtain

$$\xi_1 = \eta(A,Q,t)\,, \quad (34)$$

$$\partial C/\partial \dot{Q}\partial\,\xi_2/\partial t - \xi_2 \partial R/\partial Q = 0\,,$$
$$\xi_2 \partial C/\partial \dot{Q} = ak\exp(kt)\dot{Q}\partial C/\partial \dot{Q} + \Phi(A,t)\,, \quad (35)$$

where $\Phi(A,t)$ is an arbitrary function of A and t. Now we have

> **Theorem 1** If the infinitesimal transformations defined in eqs. (18) to (20) satisfy eqs. (34) and (35), the housing system is dynamically invariant.

Similarly, the necessary conditions for the system to satisfy divergent invariance can be expressed in the following corollary (Kobayashi et al., 1986).

> **Corollary 1** If the infinitesimal transformations defined in eqs. (18) to (20) satisfy eqs. (34) and (35), the housing system is divergently invariant.

Theorem 1 and Corollary 1 suggest that structures of the housing market with concepts of dynamic invariance and divergent invariance would be same under impacts of taste changes if the tastes changes can be described by the infinitesimal transformations satisfying Theorem 1 and Corollary 1. Necessary conditions in Theorem 1 describe the relations among the cost functions and the infinitesimal generators which must be held in order that the system is dynamically invariant. If the infinitesimal generators do not satisfy the conditions in Theorem 1, we can not use the basic model when taste changes occur. Corollary 1 tells us that while the urban area is affected by different exogenous or endogenous factors, the dynamic system is divergently invariant under the same patterns of taste changes and the same forms of cost and rent functions as those stated in Theorem 1. On the other hand, Theorem 1 and Corollary 1 imply that if we consider that structures of the housing system are described by infinitesimal transformations, and rent and cost functions, then we may say that dynamic housing systems with definitions of dynamic invariance and divergent invariance must have the same structure. Although our discussions only apply to the models built upon the PF-hypothesis, we may point out that it is possible for us to analyse other kinds of urban models, which can be kept structurally consistent under impacts of exogenous changes in a similar way. Forecast results from existing urban models are not usually accurate when exogenous taste changes alter the values of parameters and meanings of the variables in the model. However, Lie groups may supply us with a powerful tool to deal with such problems.

Now, we shall try to investigate structures of the dynamic housing system in more detail according to necessary conditions for invariance given by Theorem 1 and Corollary 1. By solving eqs. (34) and (35) in terms of infinitesimal generators, a family of infinitesimal generators which could keep the housing system invariant is obtained. If the cost functions is linear or nonlinear with respect to $\dot{Q}$ the characteristics of the partial differential equations (34) and (35) change accordingly.

(1) Linear cost function with respect to $\dot{Q}$

Let us assume that the cost function takes a form as

$$C(A,\dot{Q}) = \omega(A) + p\dot{Q}\,, \tag{36}$$

where $\omega(A)$ is an arbitrary function of A. By substituting eq. (36) into eq. (35), infinitesimal generators $\xi_2 = (\xi_{21},...,...,\xi_{2m})$ become

$$\xi_{2_j} = ak \exp(kt)\,(\Sigma_i \lambda_{ij} Q_i) + \psi_j(A,t), \qquad j=1,\ldots,m\ , \tag{37}$$

where

$$\Sigma_i \lambda_{ij} p_i = p_j\ , \qquad j=1,\ldots,m \tag{38}$$

$$\Sigma_j p_j \psi_j = \Phi(A,t) \tag{39}$$

where Φ and ψ_j are arbitrary functions satisfying eq. (39), (Kobayashi et al. 1986). ξ_1 is an arbitrary function as

$$\xi_1 = \eta(A,Q,t)\ . \tag{40}$$

(2) Non-linear cost function with respect to $\dot{Q}$

In the case of a non-linear cost function, the characteristics of the partial differential equations (34) and (35) depend upon the values of two parameters k and a. It can be proved that if neither the discount rate k nor parameter a are equal to zero, eqs. (34) and (35) have no solution. If either k or a is equal to zero. eq. (35) becomes

$$\Sigma_j \xi_{2j}\, \partial C/\partial \dot{Q}_j = \Phi(A,t)\ ,\ (=0)\ . \tag{41}$$

Only the left side of eq. (41) includes the variables Q and $\dot{Q}$. In order to hold eq. (41) for any values of Q and $\dot{Q}$, it must be held $\Phi(A,t) = 0$. The partial differential equation (41) has the following two solutions

a)
$$\begin{aligned} \tau &= a \exp(kt) + \Psi(A)\ , \\ \xi_1 &= \eta(A,Q,t)\ , \\ \xi_2 &= 0\ . \end{aligned} \tag{42}$$

b)
$$\begin{aligned} \tau &= a \exp(kt) + \Psi(A)\ , \\ \xi_1 &= \eta(A,Q,t)\ , \\ \xi_{2_j} &= \psi_j(A,Q), \qquad j=1,\ldots,m-1\ , \end{aligned}$$

$$\Sigma_j^{m-1} \psi_j(A,Q) = -\zeta(A)\, \psi_m(A,Q)\ , \tag{43}$$

where $\zeta(A)$ is an arbitrary function. Formally, we have

Theorem 2 The infinitesimal generators must satisfy eqs. (37) to (39) for the housing system to be dynamically invariant for a linear cost function with $\dot{Q}$, and the infinitesimal generators must satisfy either eq. (42) or eq. (43) for a non-linear cost function.

According to Corollary 1, we can obtain the following corollary.

Corollary 2 The infinitesimal generators must satisfy eqs. (37) to (39) for the housing system to be divergently invariant for a linear cost function with $\dot{Q}$, and the infinitesimal generators must satisfy either eq. (42) or eq. (43) for a non-linear cost function.

Theorem 2 provides the necessary conditions for the system to be invariant. These conditions determine invariant structures of the housing system under impacts of taste changes. On the basis of this theorem, we can list all of the patterns of the taste changes and forms of cost functions which make the system dynamically invariant in Table 1. Moreover, with respect to each pattern of the taste changes in Table 1, we can obtain the appropriate revenue functions from eq. (35) which make the system dynamically invariant. However, it is very difficult to obtain explicit solutions from these general conditions. In Section 8, we will solve the revenue functions according to considerations of different patterns of taste changes satisfying Theorem 2.

Table 1: Combinations of Cost Functions and Infinitesimal Transformations in Theorem 2

cost functions	a, k	infinitesimal transformations
$C(A,\dot{Q}) = \omega(A) + p\dot{Q}$ (linear)		$\tau = a\exp(kt) + \Psi(A)$ $\xi_1 = \eta(A,Q,t)$ $\xi_{2j} = ak\exp(kt)(\Sigma_i \lambda_{ij} Q_i) + \psi_j(A,t)$ $\Sigma_i \lambda_{ij} p_i = p_j \ (j=1,\ldots,m)$ $\Sigma_j \psi_j(A) = \Phi(A,t)$
$C(A,\dot{Q}) = \omega(A) + \chi(\dot{Q})$ (separably non-linear)	$a \neq 0$ $k = 0$ or $a = 0$ $k \neq 0$	$\tau = a\exp(kt) + \Psi(A)$ $\xi_1 = \eta(A,Q,t)$ $\xi_{2j} = \psi_j(A,Q)$ $\Sigma_j \psi_j(A,Q)/\psi_m(A,Q) = -\zeta(A)$
	$a \neq 0, k \neq 0$	the model is not invariant
(arbitrarily non-linear)	$a \neq 0, k = 0$ or $a = 0, k \neq 0$	$\tau = 1$ or $\tau = \Psi(A)$ $\xi_1 = \eta(A,Q,t)$ $\xi_2 = 0$
	$a \neq 0, k \neq 0$	the model is not invariant

7. CONSERVATION LAWS IN THE HOUSING MARKET

We will now use Noether's theorem to show that if the necessary conditions for the housing market to be invariant are satisfied, then there exist functions of the variables and the derivatives of the variables with respect to time which are kept constant along optimal trajectories at any point of the study period. We call the result from Noether's theorem the conservation law in the housing market. Applying Noether's theorem to the basic model, the conservation law which explains invariant relations among the variables in the housing market under the taste changes is given by

$$\Omega = (L - \dot{Q}\partial L/\partial\dot{Q}) + \xi_2 \partial L/\partial\dot{Q} = \text{const.} \tag{44}$$

For those families of infinitesimal transformations given by Theorem 2, with application of Noether's theorem we can show that the conservation laws take the following forms. For the linear function with $\dot{Q}$, the conservation law becomes

$$\Omega = (a\exp(kt) + \Psi(A))\,(U + p\dot{Q}U')\exp(-kt) + \xi_2 pU'\exp(-kt) = \text{const.} \tag{45}$$

For non-linear cost functions and the infinitesimal transformations (42), according to Noether's theorem there exists a conservation law

$$\Omega = (a\exp(kt) + \Psi(A))\,(U + U'\dot{Q}\,\partial C/\partial\dot{Q}) = \text{const.} \tag{46}$$

In the case of a non-linear cost function and infinitesimal transformations (43), we have

$$\Omega = (a\exp(kt) + \Psi(A))\,(U + U'\dot{Q}\,\partial C/\partial\dot{Q})\exp(-kt) = \text{const.} \tag{47}$$

The conservation laws explain invariant identities which must be held along the optimal trajectories under taste changes. They can explain structures in the housing system, i.e., relations among patterns of the infinitesimal transformations, values of utility, revenue and cost. Otherwise, if relations among patterns of taste changes, values of utility, revenue and cost do not satisfy the conservation law given by Noether's theorem, we can not assume that the dynamic system is invariant. This tells us that if the housing system is to be kept invariant under given patterns of taste changes, forms of utility, revenue and cost functions which can be utilized in the basic model, the model would be limited. Here it will be noted that when the functions of utility, revenue and cost keep the dynamic system invariant, it is possible to make practical tests of the conservation laws. This means that even if it is difficult to test the PF-hypothesis directly, we can do so because the conservation laws are derived from the PF-hypothesis.

8. THE STRUCTURES OF THE HOUSING SYSTEM UNDER TASTE CHANGES

By utilizing Noether's theorem, we have derived conservation laws in the housing market under different patterns of infinitesimal transformations and various forms of utility,

revenue and cost functions. Since structures of the housing market can be described by conservation laws which hold under taste changes and functions, in order to understand structures of the housing system it is very important to investigate the patterns of taste changes and the forms of utility, revenue and cost function under which conservation laws hold.

The conservation laws are partial differential equations with respect to the utility function U, the revenue function R and the cost function C. Structures of the system expressed by utility, revenue and cost functions along the optimal trajectories are defined by conservation laws (45) to (47). We shall propose several possible types of taste changes and decide the structures of the system for each. We shall also count all classes of Lie groups which make the system invariant (Kobayashi et al., 1986). In order to keep the discussion brief we shall only outline several typical cases of these patterns.

(1) Case 1 ($\tau = 1, \xi_1 = 0, \xi_2 = 0$)

In this case, the Lie groups become $\bar{t} = t + \varepsilon$, $\bar{A} = A$, $\bar{Q} = Q$. If we consider the parameter $\bar{t}$ as normal time, the infinitesimal transformations mean that changes in the tastes of households do not occur. The conservation law for the housing system can be written as

$$\Omega = (U + U'\dot{Q}\,\partial C/\partial\dot{Q}) = \text{const.} \tag{48}$$

We shall carry out our analyses when the utility functions take the different forms a) $U = \alpha + \beta F$, 2) $U = \exp(\alpha + \beta F)$, 3) $U = \ell n(\alpha + \beta F)$.

Case 1a) ($U = \alpha + \beta F$)

By substituting the utility function $U = \alpha + \beta F$ into eq. (48), we can obtain

$$\alpha + \beta F + \beta\dot{Q}\,\partial C/\partial\dot{Q} = \upsilon\,, \tag{49}$$

where υ is constant. Since $F = R - C$, we see

$$R = C - \dot{Q}\,\partial C/\partial\dot{Q} + \tilde{\alpha}\,, \tag{50}$$

where $\tilde{\alpha} = (-\alpha + \upsilon)/\beta$. If the developer is neutral to the profit, the optimal profit can be written

$$F^* = \upsilon - MC, \tag{51}$$

where $MC = \dot{Q}\,\partial C/\partial\dot{Q}$, which is the total marginal cost. In this situation, the optimal profit for the developer is equal to υ - MC at any time.

Case 1b) ($U = \exp(\alpha + \beta F)$)

The conservation law in this case is

$$\Omega = \exp(\alpha + \beta F)(1 + \beta\dot{Q}\,\partial C/\partial\dot{Q}) = \upsilon\,. \tag{52}$$

We can obtain

$$R = C + 1/\beta\, \ell n(\upsilon/(\beta\, MC + 1)) - \alpha/\beta\,. \tag{53}$$

The optimal profit is

$$F^* = 1/\beta\, \ell n(\upsilon/(\beta\, MC + 1)) - \alpha/\beta\,. \tag{54}$$

Case 1c) $(U = \ell n(\alpha + \beta F)$.

In this case, the conservation law is expressed as

$$\ell n(\alpha + \beta F) + \beta/(\alpha + \beta F)\dot{Q}\, \partial C/\partial \dot{Q} = \upsilon\,. \tag{55}$$

It can be seen that

$$((\alpha + \beta R - \beta C)\exp(-\upsilon))^{(\alpha + \beta R - \beta C)} = \exp(-\beta MC)\,. \tag{56}$$

Eq. (56) is an implicit function of the optimal profit. It is impossible to make an explicit expression of the optimal profit. That is, if we employ a logarithmic utility function, we can not explicitly express the optimal profit which makes the system invariant under taste changes.

Case 2 ($\tau = 1$, Linear cost function)

In this case, the cost function takes the form of

$$C(A,\dot{Q}) = \omega(A) + \Sigma_j\, p_j\dot{Q}_j\,. \tag{57}$$

According to Theorem 2, in order for the system to be invariant the infinitesimal transformations have to take the form of

$$\tau = 1\,, \tag{58}$$

$$\xi_1 = \eta(A,Q,t)\,, \tag{59}$$

$$\xi_{2_j} = \psi_j(A)\,,\ j=1,\ldots,m-1 \tag{60}$$

$$\Sigma_j\, p_j\psi_j(A) = \Phi(A)\,. \tag{61}$$

If eqs. (58) to (61) hold, the conservation law takes the form of

$$\Omega = (U + p\dot{Q}U' - \Phi(A)U') = \upsilon(t)\,, \tag{62}$$

where $\upsilon(t) = \upsilon \exp(kt)$. By solving eq. (35), the revenue function which keeps the basic model invariant is

$$R(A,Q) = \theta_1(A) + \theta_2(\chi\pi(A)) , \tag{63}$$

where

$$\chi = \Sigma_j Q_j + \zeta(A)Q_n , \tag{64}$$

$$\zeta(A) = - \Sigma_j \psi_j (A)/\psi_m (A) , \tag{65}$$

in which θ_1, θ_2, π, and ζ are arbitrary functions.

In eq. (59), since the infinitesimal generator ξ_1 for tastes of the amenities is an arbitrary function with respect to amenities, service values and time, we can say that taste changes about amenities have no direct impact upon invariance of the dynamic system. From eq. (60), we see that the tastes of the households for the service values are dependent only upon the amenities. It should be noted that if we make certain assumptions about properties of the revenue functions in eq. (63), we can obtain practically testable revenue functions. For instance, (1) if we consider that $\pi(A)$ in eq. (63) is constant, the revenue function becomes separable with respect to A and Q; (2) if we consider $\theta_1(A)$ to be constant, we obtain a multiplicative revenue function.

In Case (1), the basic model is the same as the existing models that are based upon the PF-hypothesis (Fujita, 1983). In order words, most of the existing models based upon the PF-hypothesis can be considered as a special case of our model, where the revenue functions satisfy the necessary conditions for system to be invariant. In Case (2), since $\zeta(A) = 1$, the variable Q in the revenue function is not changeable with respect to $\zeta(A)$; however, the patterns of the taste changes are limited by $\Sigma_j \psi_j (A) = 0$.

Case 2a) $(U = \alpha + \beta F)$

In this case, the conservation law is

$$\alpha + \beta F + \beta(p\dot{Q} - \Phi(A)) = \upsilon(t) , \tag{66}$$

where $\upsilon(t) = \upsilon \exp(kt)$. The optimal profit is

$$F^* = (\upsilon(t) - \alpha)/\beta + \Phi(A) - MC , \tag{67}$$

where $MC = p\dot{Q}$, $\upsilon(t) = \upsilon \exp(kt)$.

Case 2b) $(U = \exp(\alpha + \beta F))$

In this case, the conservation law and the optimal profit can be expressed

$$\exp(\alpha + \beta F)(1 + \beta MC - \beta \Phi(A)) = \upsilon(t) , \tag{68}$$

$$F^* = \ell n(\upsilon(t)/\beta MC - \beta \Phi(A) + 1))/\beta - \alpha/\beta . \tag{69}$$

Case 2c) $(U = \ell n(\alpha + \beta F))$

Just as in Case 1c) the optimal revenue can not be explicitly expressed in this case.

From eq. (61), we can interpret $\Phi(A)$ as the value measured by the marginal cost that is required for improvement of housing services. The need for improvement comes from changes in the tastes of the households. Now, consider the case when the tastes of the households for housing services are increased. In this case, since $\Phi(A)$ is greater than

zero, from eqs. (67) and (69) we can see that the developer can obtain the subsidized profit $\Phi(A)$ resulting from taste changes. The subsidized profit is only related to amenities. That is, if there does not exist scale effects in housing production, under various location conditions the patterns of taste changes and also the subsidized profits should be different.

Case 3) ($\tau = 1$, Nonlinear cost function)

Now we shall take scale factors in the construction of houses into account. According to Theorem 2, if the housing system is invariant, the cost function has to take the separable form

$$C(A,\dot{Q}) = \omega(A) + \chi(\dot{Q}) , \tag{70}$$

and the infinitesimal transformations satisfy

$$\tau = 1 , \tag{71}$$

$$\xi_1 = \eta(A,Q,t) , \tag{72}$$

$$\xi_{2_j} = \psi_j(A,Q) , \ (j=1,...,m-1) , \tag{73}$$

$$\Sigma_j \psi_j (A,Q)/\psi_m(A,Q) = - \zeta(A) . \tag{74}$$

This case is different from Case 2 in that the invariance of the housing system can be held even if the tastes of the households are explicitly related to levels of housing services. The functions $\psi_j(A,Q)$ in eq. (73) may take separated forms as $\psi_j(A,Q) = \psi_j(A)\kappa(Q)$. If we consider that $\kappa(Q)$ is constant, then the infinitesimal transformations are the same as those in Case 2. From these discussions, we conclude that there are wider families of taste changes here than in Case 2 which make the housing system invariant. Otherwise, revenue functions in eqs. (63) to (65) also make the system invariant under the patterns of taste changes in this case. From the results above, we see that from a practical viewpoint types of revenue functions would multiplicative or separable with respect to A and Q.

Case 3a) ($U = \alpha + \beta F$)

If we let $MC = \dot{Q}\,\partial C/\partial \dot{Q}$ the conservation law and optimal profit are respectively

$$\alpha + \beta F + \beta MC = \upsilon(t) , \tag{75}$$

$$F^* = \upsilon(t)/\beta - MC - \alpha/\beta . \tag{76}$$

From eq. (76) it can be seen that the scale factor increases the developer's profit.

Case 3b) ($U = \exp(\alpha + \beta F)$)

The conservation law and the optimal profit are given by

$$\exp(\alpha + \beta F)(1 + \beta MC) = \upsilon(t) , \tag{77}$$

$$F^* = \ell n(\upsilon(t)/(\beta MC + 1))/\beta - \alpha/\beta \,. \qquad (78)$$

Case 3c) $(U = \ell n(\alpha + \beta F))$

Just as in Case 1c) the optimal profit can not be explicitly expressed here.

9. CONCLUSIONS

Upon the basis of the assumption of the developer as a profit-taker and by the use of infinitesimal transformations in Lie group theory, we have investigated the behavior of the developer in the housing market under impacts of taste changes of households. This study applies Lie group theory to regional science and urban economics. Although it is far from a comprehensive theory we believe that we have demonstrated the potential applicability of Lie groups in the theoretical analysis of the behavior of developers. In fact, Lie group theory can serve as a powerful "tool" to capture taste changes and to enable us to investigate the behavior of human beings in this situation.

From this study we have obtained the following results: (1) there does not exist any model for the behavior of developers built upon the PF-hypothesis which is invariant under arbitrary patterns of the taste changes of households; (2) if the revenue function takes a multiplicative or separable form with respect to A and Q, the basic model is invariant under certain patterns of taste changes; (3) some existing models built upon the PF-hypothesis for the behavior of developers in the housing market agree with our model, which is invariant under certain patterns of taste changes; (4) if the dynamic model is invariant under taste changes, then there exist conservation laws in the housing market which are observable; (5) if developers in the housing market are perfectly competitive and there is no scale factor on construction costs, the developer can obtain "subsidized" profit from changes in the tastes of households, (6) if there exist scale factors in construction costs, the "subsidized" profit of the developer can be obtained from the scale factors and changes in the tastes of households.

ACKNOWLEDGEMENTS

The authors would like to express thanks to Professors Å.E. Andersson, B. Johansson, G. Haag and M. Fujita who have provided us with very valuable suggestions for this study.

REFERENCES

Allen, P.M., N. Sanglier, G. Engelen, and F. Boon, 1985, "Towards a New Synthesis in the Modelling of Evolving Complex Systems", *Environment and Planning A*, 15:543-550.

Barentsen, W., and P. Nijkamp, "Modelling Non-linear Processes in Time and Space", Discussion Paper presented to the Regional Science Association European Advanced Summer Institute, 1986.

Diamond, B.D. et al., 1982, *The Economics of Urban Amenities*, Academic Press, New York.

Fujita, 1983, Urban Spatial Dynamics: A Review, *Systemi Urbani*, 3: 411-475.

Ikeda, M. et al., 1975, "On the Concept of Symmetry in Pontryagin's Maximum Principle, *SIAM. Jour. of Control*, 13:389-399.

Kobayashi, K., W.B. Zhang, and K. Yoshikawa, 1986, "Behavior of Developers and Conservation Laws in the Housing Market", Kyoto Univ., Research Report, 86-PT-2.

Samuelson, P.A., 1976, "Speeding up of Time with Age in Recognition of Life as Fleeting", in A.M. Tang et al., (eds.), *Evolution, Welfare, and Time in Economics, Essays in Honor of Nicolis Georgescu-Roegen*, Lexington Books, Lexington, Massachusetts.

Sato, R., 1981, *Theory of Technical Change and Economic Invariance - Application of Lie Groups*, Academic Press, New York.

Wilson, A.G., 1981, *Catastrophe Theory and Bifurcation, Application to Urban and Regional Systems*, Croom Helm, London.

LIST OF CONTRIBUTORS

Åke Andersson
Department of Economics
University of Umeå
S-901 87 Umeå
Sweden

Jean-Pierre Aubin
CEREMADE
Université de Paris-Dauphine
F-75775 Paris cx(16)
France

Wim Barentsen
Department of Economics
Free University
P.O.Box 7161
NL-1007mc Amsterdam
The Netherlands

David Batten
CERUM
University of Umeå
S-901 87 Umeå
Sweden

Gerald Carlino
Federal Reserve Bank of Philadelphia
6th and Arch Streets
Philadelphia, Pennsylvania 19106
USA

Lata Chatterjee
Department of Geography
Boston University
48 Cummington Street
Boston, Massachusetts 02215
USA

Bruno Dejon
Institute for Applied Mathematics
University of Erlangen-Nürnberg
Martensstrasse 3
D-8520 Erlangen
Federal Republic of Germany

Bernhard Güldner
Siemens AG
UB Energie- und Automatisierungs-
teknik
Gleiwitzer Strasse 555
D-8500 Nürnberg
Federal Republic of Germany

Günter Haag
Institute for Theoretical Physics
University of Stuttgart
Pfaffenwaldring 57/III
D-7000 Stuttgart 80
Federal Republic of Germany

John Hartwick
Department of Economics
Queen's University
Kingston K7L3N6
Canada

Börje Johansson
CERUM
University of Umeå
S-901 87 Umeå
Sweden

Kiyoshi Kobayashi
Department of Social Systems
Engineering
Tottori University
Tottori 680
Japan

Robert Kuenne
Department of Economics
Princeton University
Princeton
New Jersey 08544
USA

T.R. Lakshmanan
Department of Geography
Boston University
48 Cummington Street
Boston, Massachusetts 02215
USA

Edwin Mills
Department of Economics
Northwestern University
Evanston, IL 60201
USA

Anna Nagurney
School of Management
University of Massachusetts
Amherst, Massachusetts 01003
USA

Peter Nijkamp
Department of Economics
Free University
P.O.Box 7161
NL-1007mc Amsterdam
The Netherlands

Tönu Puu
Department of Economics
University of Umeå
S-901 87 Umeå
Sweden

John Quigley
Graduate School of Public Policy
University of California, Berkeley
2607 Hearst Avenue
Berkeley, California 94720
USA

Hans-Holger Rogner
IIASA
A-2361 Laxenburg
Austria

Tony Smith
Regional Science Department
University of Pennsylvania
Philadelphia PA 19104
USA

Michael Spencer
Department of Economics
Queen's University
Kingston K7L3N6
Canada

Pravin Varaiya
College of Engineering
University of California, Berkeley
2607 Hearst Avenue
Berkeley, California 94720
USA

Georg Wenzel
TE KA DE
Thurn-und-Taxis Strasse 14
D-8500 Nürnberg
Federal Republic of Germany

Kazuhiro Yoshikawa
School of Civil Engineering
Kyoto University
Kyoto 606
Japan

Wei Bin Zhang
CERUM
University of Umeå
S-901 87 Umeå
Sweden

Index

ISBN: 0 444 87357 0