PSEUDO-DYNAMIC URBAN MODELS J. M. BATTY Ph.D. 1984

# PSEUDO-DYNAMIC URBAN MODELS.

by

JOHN MICHAEL BATTY.

Professor of Town Planning,
University of Wales Institute of Science and Technology.

A Thesis submitted for the Degree of Doctor of Philosophy of the University of Wales, 1984.

# SUMMARY.

This thesis explores the mathematical properties of urban equilibrium models in terms of their structural forms and solution procedures. The class of models investigated is based on linear input-output relationships and nonlinear spatial interaction principles, which are elaborated as spatial multipliers and have a real-time dynamic interpretation. These models are usually solved using iterative methods which also involve a pseudo- or solution-time dynamics. The central task of this thesis is to match these real and pseudo-time dynamics and to exploit their parallelism in efficient model solution and the generation of new model structures.

The class of models is first generalised through linear equilibrium and nonlinear optimisation theory, thus setting the context for an elaboration of their dynamic properties. A fully-dynamic structure is derived and then collapsed back to pseudo-dynamic form in which both static and dynamic components exist. A typology of pseudo-dynamic models is derived, and the notion of enabling efficient model solution through pseudo-dynamics is demonstrated, first for problems involving locational constraints, then for adaptive calibration of the model's spatial interaction parameters. A fully integrated solution procedure is then developed embodying matrix iterative analysis and analogies with control theory.

A more traditional mode of analysis is also presented in treating the spatial

interaction dynamics as a Markov process. Results from discrete Markov chain theory are applied, thus enabling the sensitivity and spatial invariance between model inputs and outputs to be assessed. These methods which involve the way such models are structured, cast light on the empirical quality of many previous applications using these types of model. This theory is generalised to models with many inputs and an empirical demonstration provided. Throughout the thesis, models are tested using data from the towns of Reading, Peterborough, Greater London, Central Berkshire and Melbourne.

STATEMENT AND DECLARATION.

The research reported in this thesis is the result of my own investigations.

JOHN MICHAEL BATTY.

DIRECTOR OF STUDIES.

This work has not already been accepted for any degree and is not being concurrently submitted for any degree.

JOHN MICHAEL BATTY.

# PREFACE.

I have been associated with the urban modelling field since its early days in the 1960's. As a junior staff member in the University of Manchester, I was attracted to the area when it was widely thought that mathematical social science held the key to the future. Nothing I have come across since then has dissuaded me from that view although the view is no longer popular among less mathematically inclined social scientists. This thesis represents some of my more recent work in this field which builds on some of the earliest developments. During the last 20 years, I have had the good fortune and privilege to meet most of the contributors to this field and I have collaborated with many of them. It might therefore seem strange to some that I should be submitting a doctoral thesis when I have researched the area for so long. A word of explanation is thus warranted, and in any case, this preface also provides me with the all-too-rare opportunity to indulge in some personal reminiscences which provide a context.

The material I have included in this thesis represents a rather coherent theme in urban modelling research. It is concerned with technical questions of model solution which have both substantive implications for urban systems theory and implications for the relevance of empirical work. About one third of the work has already been published in journal article form or in conference proceedings, but the rest has not. Although the material hangs together well and represents a line of research I have worked on for some

8 years, I do not feel that all this material can be published in journal article or book form. In a sense, it is *too* technical. That is to say, some of the material is so detailed that journals in the fields which specialise in this research have not really developed to the point where the publication of such detailed work is appropriate. In short the lack of any normal science in this field is reflected in the types of work journals feel able to publish. I make this point rather strongly in both the Introduction and Conclusions to the thesis, and readers will note its recurrence elsewhere. In this sense then, a thesis is an ideal place to publish such work in an integrated way which should appeal to a small group of dedicated researchers. This is the main reason why I have put the material together in this form.

I have many people to thank, for helping me in this research. From Manchester days, George Chadwick set me on the modelling track which took me to Reading where I enjoyed a fruitful cooperation with Dave Foot and Peter Hall. But it was in the Faculty of Engineering in the University of Waterloo where these ideas firmly took root in my sojourn there in 1974-75. In Waterloo, Lionel March and myself began to explore sequential processes in urban models using information-minimising principles and we published two papers (Batty and March, 1976; Batty and March, 1978) which form the natural antecedents of this research and are strongly reflected in Chapters 3 to 6. I cannot now remember how I got started on pseudo-dynamic models as such, but in late 1976, it was Pedro Geraldes who told me of similar work by Yossi Berechman. I wrote to Berechman and we met in Buffalo in the summer of 1977. During this period, I read Ian William's papers on the subject and we also corresponded in 1978. That was the period when I wrote early versions of Chapters 3 to 7.

My work really received a boost from the publication of Arie Schinnar's paper in 1978 which led to Chapter 10 and to my interest in Markov processes in urban models. Then at UWIST in 1981, I began to work on James Coleman's Model of Collective Action (Coleman, 1973) in an entirely different context. Somehow, Coleman's model provided a more general framework for conventional urban models in one sense and a more specific one in another. Chapter 11 was the result of these ponderings and was researched and written at the University of Melbourne where I was a Visiting Fellow in 1982. Richard Spooner helped me a lot with these ideas and I have to thank the Economic and Social Research Council (ESRC, formerly SSRC) for grant support which enabled me to write Chapters 2 and 11, and to employ Richard Spooner.

I have to thank my secretary, Beryl Collins, who has done such a magnificent job typing the thesis at the same time as helping me in all my other diverse tasks. Finally, I thank my family for their tolerance of my indulgence in what to them must seem the most curious of projects.

MICHAEL BATTY
Welsh Saint Donats,
Vale of Glamorgan.

March 1984.

# ACKNOWLEDGEMENTS.

Chapter 2 was first presented to the *International Symposium on New Directions in Urban Modelling*, July 11-15, 1983, at the University of Waterloo, Ontario, and is forthcoming in the proceedings of that conference. Chapter 8 was first presented at the *International Research Conference on Spatial Interaction Theory and Planning Models*, August 29-31, 1977 at Bastad, Sweden, and is published in A. Karlquist, L. Lundquist, F. Snickars and J. Weibull (1978) *Spatial Interaction Theory and Planning Models* (North-Holland Publishing Company, Amsterdam, pp. 227-252). Chapter 9 was presented to the 10th Annual Conference of the British Regional Science Association, at University College, London, September 3rd, 1977 and is published in *London Papers in Regional Science, 9*, 26-63, 1979. Chapter 10 is published in *Environment and Planning A, 11*, 487-497, 1979. Chapter 11 was first presented at the 29th North American Regional Science Association Meetings, November 12-14, 1982 in Pittsburgh, Pennsylvania, and is forthcoming in *Papers of the Regional Science Association*, Volume 52.

CONTENTS.

Page

## LIST OF FIGURES.

Page

## LIST OF TABLES.

# CHAPTER 1.

# INTRODUCTION.

> "The main role of models is not so much to explain or predict - though ultimately these are the main functions of science - as to polarize thinking and to pose sharp questions."
>
> Mark Kac, *Science, 166,* 1969, p.699.

The purpose of this thesis is to pose sharp questions, questions concerning the potential for improving conventional urban models to the point where such models become operationally useful, and questions concerning the conditions under which such models can be used in the most constructive way. It is my belief that there is considerable latent potential within conventional urban models for addressing a variety of urban problems and that this potential has, by and large, not been realised so far. Although the models which form the starting point of this thesis have been developed throughout the last 20 years, there have been so many different approaches and so few researchers, that most of the field has been preoccupied with examining rather dramatically different strategies for model use and design. Consequently, the more painstaking, laborious 'normal' science which follows in the wake of new approaches, has not been engendered in this field. Urban modelling like many areas in the social sciences, thus lacks the development of a 'normal' science in Kuhn's (1970) terms.

In the social sciences, various commentators explain the preoccupation

with conflicting paradigms, the lack of consensus over any 'correct' approach, and the inability to develop highly contextual detailed research in various ways, but generally either as evidence of a pre-scientific situation, or as an intrinsic feature of social science theory and knowledge. However, there do exist areas of the social sciences such as mathematical economics and psychology which appear to be characterised by normal scientific activity, notwithstanding the considerable controversy which surrounds the value of such work. Urban modelling is similar to these areas and the fact that it has not been characterised so far by normal scientific activity is due to the small size of its research effort and its strong links to practice, particularly planning practice, which has been particularly unstable over the last two decades.

It is the contention of this thesis that a major reevaluation of what has been developed in urban modelling by researching the detail of conventional models, will lead to new insights of profound importance to the field. Rather than changing one's approach to modelling when models do not seem to be yielding the desired results, it might be possible to modify existing models to cope with such problems, but only after much more detailed research into their structures has been accomplished. There is a dramatic example of this strategy in this thesis. In the evolution of the field, it was thought that different approaches would be required to resolve the inadequacies of conventional urban models in their treatment of time, their ability to characterise appropriate economic processes and such-like. The field is characterised by such shifts but failure to explore existing models to an appropriate level of detail has not enabled the importance of model

structure in generating results to be determined.

One conclusion of this thesis is that conventional models may produce trivial spatial results because the model's mechanisms are spatially insensitive to the way its structure is designed. This finding puts in doubt many model applications made during the last 20 years. It is a conclusion which takes most of this thesis to develop and it is built on a detailed knowledge of how the field has developed. But the fact that it has not been picked up by the field as a major question to explore shows the field's preoccupation with 'big questions' which spinoff from the obvious, superficial limitations of urban models, rather than the small, not so obvious, indeed hidden limits which only indirectly reveal 'big questions'. This is a negative conclusion of this thesis but equally there are many positive findings which aid better solution and design of conventional model structures.

## TWO DECADES OF URBAN MODELLING.

The first urban models were built in the United States between 1960 and 1965, and during that time, four very different approaches to simulating urban land use and activity systems emerged. These were embodied in four very different modelling styles. The simplest based on linear representations of known urban relationships and behaviour is best seen in econometric models such as the EMPIRIC (Hill, 1965). Such models unlike their economic counterparts, were not based on well-tempered urban theory but on commonsense relationships. In contrast, their solution was by the latest techniques of econometric analysis. The second approach was also based on known

urban relationships but embodied these in nonlinear ways, and particularly made use of well-known nonlinear gravitational relationships. The Lowry model (Lowry, 1964) is the most famous example, based on more considered urban theory than the EMPIRIC model but solved in a more *ad hoc* manner.

These two model types were based on known relationships and did not contain any implication that such behaviour could be explained as the outcome of an optimising process. In contrast, both linear and eventually nonlinear model types emerged where optimisation constituted the basic way of representing urban relationships. The emphasis on using urban models in plan-making led to linear optimising models based on linear programs which 'explained' or 'optimised' the distribution of land uses and activities through minimisation of some cost function. Schlager's (1965) Land Use Plan Design Model is the clearest example. Other models based on individual optimising strategies as contained in the rationality assumptions of utility theory were developed. The model due to Herbert and Stevens (1960) was a linear programming approximation of Alonso's (1964) theory of the urban land market based on micro-economic utility theory.

These distinctions - linear versus nonlinear, and optimising versus nonoptimising - reflect basic dimensions which continue to characterise urban models, but since the early 1960's, there has been a considerable effort to unify these differences in the effort to see models as special cases of a more general model structure. As Chapter 2 demonstrates, this quest has been extremely successful and the field is now characterised by a much clearer view of how these different

model types relate to one another. Moreover, this unification has succeeded in showing how spatial interaction is central to land use - activity modelling, and how macro (regional) and micro (urban) economic theory can be linked to statistical optimisation and econometric analysis through disaggregate theory.

Many of the early reviews of urban modelling (see for example, Lowry, 1965) identified another distinction involving the treatment of time. Models were classified as static or dynamic with the assumption that static models represented simplifications, aggregations or cross-sections of models based on dynamic processes. However during the last 20 years, progress in building dynamic urban models has been slow. Many of the earlier attempts involved simply indexing what were in effect static models, in terms of time, and thus contained no theory or analysis of dynamic behaviour. In Samuelson's (1948) famous phrase, time was *not* involved in an 'essential way' in such structures.

It was not until Forrester (1969) published *Urban Dynamics* that attention became focussed on dynamic processes *per se* and even then, Forrester's approach emphasised issues involving simulation and complexity over and above dynamics. In the last five years, however, an entirely different approach to dynamics has emerged based on embedding static urban models into dynamic processes; that is, by coupling existing urban models reflecting equilibrium conditions to processes generating disequilibrium, such processes being based on the dynamics of discontinuity, catastrophe and fluctuation. Wilson's (1981) work is central to this as is Allen, Sanglier, Boon, Deneuborg

and De Palma's (1981) work and operational models embodying similar ideas are already making their appearance (Schneider, 1976; Varaprasad and Cordey-Hayes, 1982).

These developments in linear modelling, optimisation and dynamics all constitute an essential backcloth against which this thesis has been written. In particular, the starting point of this thesis will be in terms of linear modelling, but in developing a linear analysis of existing model structures, forays into optimisation theory particularly relating to the statistical derivation and calibration of spatial interaction models, will be made. The recent advances in urban model dynamics just referred to will not constitute a theme to be developed here. They do however provide an important contrast with the dynamics presented here in that the dynamics elaborated in the sequel are concerned with the mechanisms and behaviour of static model solution, rather than the wider processes of urban dynamics. Indeed, the reference to pseudo-dynamics in the title to this thesis reflects a definition of the dynamics of model solution, not the dynamics of urban processes.

To set the context to this work, the second chapter will present a review of developments in linear modelling and optimisation theory. These developments are those concerned with the unification of the field referred to above and in particular, Chapter 2 will emphasise the value of linear analysis in guiding research into model structure. In essence, this thesis takes as its starting point the traditional static urban model such as that portrayed by Lowry (1964) but as Chapter 2 emphasises, this model is one from a very general class of

urban models. It is thus essential to continually generalise the results from this thesis to this broader class.

The models presented here will be developed in linear form as equilibrium conditions and the general goal of this thesis is to examine the kinds of linear dynamics which give rise to such equilibria. These linear structures will initially be torn apart and their cross-sectional mechanisms elaborated into fully-dynamic structures which will then be collapsed back to cross-sectional form. But on the way, the idea of a pseudo-dynamic model will be identified as a structure in which its explicit dynamics aids the solution of its cross-sectional form through notions concerning iteration and the reallocation of activity.

Once the pseudo-dynamic theory has been developed, it is then used to examine three problems involved in solving conventional urban models: the incorporation of locational constraints, the calibration of the model's global parameters on spatial interaction, and the simultaneous solution of the model to incorporate constraints and to calibrate parameter values. Various algorithms associated with these mechanisms are tested, with each algorithm introduced representing an improvement or generalisation of the preceding one. However out of this research there then comes a theoretical statement of pseudo-dynamic processes based on the distinction between multiplier and spatial Markov processes. As is indicated in Chapter 2, these various developments present aids to model design and development in an empirical context which lead to important results concerning model calibration and spatial variation.

THE CHRONOLOGY AND ORGANISATION OF RESEARCH.

The ideas on which this thesis is built relate to research undertaken by the author beginning over a decade or so ago. In particular, the basic idea of tearing apart the linear structure of the Lowry model and elaborating it in dynamic form through its multiplier relationship, was used as the basis for a fully-dynamic urban model of the Reading region in 1971 (Batty, 1976). This type of urban model was also subsequently developed by Ayeni (1979) and by Webber (1979). A parallel stream of research by the author relates to examining the dynamic implications of spatial interaction modelling and in particular, a dynamic form of information-minimising developed as a two-stage process (Batty and March, 1976) and as a temporal form (Batty and March, 1978). These ideas were also developed simultaneously by Snickars and Weibull (1977) and later by Webber (1979), and this has important implications for the way spatial interaction models are handled in Chapters 3 to 9.

Other research has been influential in guiding the work reported here. For example, the pseudo-dynamic form of model developed in Chapters 3 to 9 has similarities to Berechman's (1976) work. The complete movers pseudo-dynamic model presented first in Chapter 4 but exhaustively in Chapters 8 and 9 is closely related to Baxter and Williams' (1975) model. The motivation for Chapters 10 and 11 is based on relating the ideas of Chapters 6 and 7 to Schinnar's (1978) work on spatial invariance in the Lowry model. Yossi Berechman, Ian Williams and Arie Schinnar all commented on the relationship of these ideas to their own when they were first explored by the author. Since then, the ideas of

Chapters 10 and 11 (and of 2) have been related to other linear models, particularly to Coleman's (1973) model of collective action based on social exchange, and the author has developed these ideas in rather different vein (Batty, 1981a).

It is extremely important for the reader to be aware of the steps in this research as contained in the subsequent chapters for it is all too easy to disguise one's tracks in an effort to present a finished product. Chapters 3 to 11 were written and researched over a five year period from 1977 to 1982 in the given order, and Chapter 2, the review, was written last. In fact, Chapters 3 to 9 were researched over an intensive period in 1977-1978 while Chapter 10 was written in 1979. At this point, the ideas of Chapter 10 were developed in a different context as generalisations of Coleman's (1973) model, and as this work accomplished in 1980-1981 is not strictly concerning urban models in the sense portrayed here, it has not been included in this thesis. Finally, this work veered back to urban modelling and Chapters 11 and 2 were written in 1982 and 1983 respectively.

A number of case studies have been employed here based on data for the Reading urban area (1966), Peterborough New Town (1971), the London Traffic Study Region (1964), the Area 8 (Central Berkshire) Planning Region (1971) and Greater Melbourne (1976). As these data bases are not developed substantively here (the emphasis in this thesis is not on the case studies *per se*), no details are given although the interested reader is referred to other articles by the author for such details (Batty, Bourke, Cormode and Anderson - Nicholls, 1974; Batty, 1976; Batty, 1978).

Finally, the chapters are organised as follows. Chapter 2 sets the context to this research by a review of linear model structures and optimisation theory, but as this chapter contains a synthesis, several new results particularly relating to input-output analysis and linear urban models, are presented. In Chapter 3, the idea of a pseudo-dynamic model is developed through a fully-fledged dynamic model incorporating movers and stayers which is collapsed back to closed form. A typology of model types is generated from this closed form in Chapter 4, and one model type from the typology adapted to handle locational constraints in Chapter 5.

In Chapter 6, another version is elaborated to handle locational constraints and to calibrate spatial interaction parameters, and an algorithm for this is developed in Chapter 7. In Chapter 8, a further model based on complete mover streams of redistribution and relocation, related to matrix iterative analysis is presented, and algorithms for this developed in Chapter 9. The thesis then changes direction and introduces Schinnar's (1978) results in Chapter 10 where these are generalised in terms of the Markov representation of spatial averaging contained in Chapter 6. Finally, this generalisation is taken further and given empirical support using data from Melbourne in Chapter 11. Conclusions are then briefly drawn out in Chapter 12.

NOTATION AND PRESENTATION.

Each chapter is relatively self-contained in that ideas and equation systems introduced in earlier chapters are repeated if they are to be elaborated in the given chapter. This duplication is quite purposeful

and never excessive, and it is also essential in that as the notational requirements of different parts of this thesis vary, it is necessary to redefine notation occasionally. Moreover, this enables key ideas to be continually emphasised. At times, repetition is necessary because a model is presented with a slightly different emphasis. For example, the Lowry model in Chapters 2, 10 and 11 is presented in distributional terms in contrast to the same type of model in Chapters 3 to 9 which is given in absolute activity terms.

Throughout the text, an effort has been made to maintain notational consistency. For example, zones are normally subscripted by i and j, although the number of zones is defined variously as N, M, n, m or I, J. Other indices k, l, m, n are used more generally for activities and/or zones. In terms of the time dimension, time is indexed by t, $\tau$, T, sometimes by m, n. Where possible variables are defined using an obvious notation: for example, E, e as employment, P, p as population, but this depends on context. Where population is endogenous, employment exogenous, y and x respectively are used to emphasise the causal relationship.

Vectors and matrices are defined by underlining. Lower case underlines, e.g.: $\underline{x}$ are 1xN row vectors while upper case e.g.: $\underline{A}$ are NxN matrices. Matrix multiplication is also based on consistent dimensioning. Column vectors are usually transposed row vectors, for example $\underline{x}'$ or $\underline{x}^T$ but note that the prime' can also mean the first derivative or a general index while T can be trips or time. The precise meaning will be obvious from the context. In the sequel, models will normally be represented in matrix terms where a row vector

$\underline{y}$ is output from a row input vector $\underline{x}$ transformed by $\underline{A}$ as $\underline{y} = \underline{x}\,\underline{A}$ However in Chapter 10, to preserve the comparison with Schinnar's (1978) paper, such models are given as $\underline{y} = \underline{A}\,\underline{x}$ where $\underline{y}$ and $\underline{x}$ are now column vectors, with $\underline{y}$, $\underline{x}$ and $\underline{A}$ clearly transposes of the usual row defined variables. Apart from these differences, this thesis is presented using the conventional notation of the field as seen in books such as Wilson (1974), Batty (1976), Oppenheim (1980) and Foot (1981).

# CHAPTER 2.

## LINEAR AND NONLINEAR STRUCTURES FOR URBAN MODELS.

In a field as rich and diverse as urban modelling, it is tempting to begin a review of technical developments during the last twenty years by attempting as wide a synthesis as possible. However during its short history, the field has been characterised by the development of certain significant themes and in this introductory chapter, it is proposed to review a limited number of such themes to give a flavour of the achievements and difficulties which characterise the wider field. Moreover, a technical review such as this one must seek to synthesise developments in a constructive way for the main purpose of this chapter is to point the way forward within the limits set rather than just review the past. Thus a major conclusion from this review will relate to those research questions which emerge from the present state-of-the-art, some of which will form the themes to be developed during this thesis.

It is useful to first characterise the development of urban modelling in terms of theoretical contributions and practical applications and elaborations. When the field first emerged in the early 1960's, developments were practice-led and rooted in empirically defined problems. Since then the practical

context to such work has changed dramatically, and together with the gradual maturation of the field, modellers have looked harder and deeper into the theoretical foundations of urban systems. The field is now quite different from its early form in that theoretical work now dominates and there is a comparative dearth of practical applications. However its development is characterised by one central theme which relates to the unification of a variety of modelling styles and techniques, and it is now possible to tie together the diversity which characterises the field in a way which has only become possible quite recently.

Unification is best seen in the way descriptive and predictive models of urban structure can now be tied to their prescriptive counterparts. The main way in which such 'behavioural' models have been linked to 'normative' models is through ideas about optimisation, in terms of substantive questions related to optimising behaviour and through optimisation methods. It is now possible to see the models of the 1960's such as those associated with Lowry (1964), Herbert and Stevens (1960) and Schlager (1965) as forming part of a more general model framework in which each can be regarded as a particular case. Closely connected to such developments is the notion that realistic model structures are neither linear nor nonlinear in terms of the way variables are related but that both forms of technique can and should be used to illuminate model structure. This intermixing of different modelling styles appears in many guises: in exploring models as accounting structures, in model calibration, in extending models to deal with multi-activities and in integrating major sectors of the urban and regional system such as the demographic and economic sectors. The major achievement of the last decade has in fact been in enabling such diversity to be explained, elaborated and developed in a unified framework, and one major goal of this review is to illustrate how this has and is being pursued.

Several other themes can be considered significant. The essential unity of land use and transportation was the watchword of the 1950's although in the early development of urban modelling, many models were designed in which transportation was implicit or even absent. Since then relevant models of macro-urban or metropolitan structure have always been underpinned by ideas concerning spatial interaction and this may partly explain the dearth of econometric modelling at the urban land use level. The field has also been preoccupied by questions of statics and dynamics. Static models have come to dominate until comparatively recently for most developments of dynamic models have been static in their original conception. Less progress has been made here than in the area of optimisation although of late new conceptual insights into the evolution of urban systems have been generated by studies of theoretical urban dynamics. In particular, questions of disequilibrium in terms of demand and supply issues, and notions concerning discontinuity and threshold have been addressed through dynamics, and there have been attempts to treat static models as the equilibrium conditions of such dynamic processes (Allen, Sanglier, Boon, Deneunbourg and DePalma, 1981; Wilson, 1982). In this thesis, some of these ideas will be hinted at through concepts involving the dynamics of model solution rather than urban dynamics.

The broader, more substantive issues will not be addressed here. Suffice it to say that there are still major questions concerning the relevance of those aspects of the urban system which are embodied in the mathematical urban models treated here. Although modelling techniques have advanced dramatically in the last twenty years, the system being modelled has remained reasonably stable, notwithstanding equally dramatic changes in the perceptions of what planners and policy-makers consider to be significant in urban terms. Although these questions are important, this review will

focus on those narrower technical developments associated with the unification of the field alluded to above.

As an introduction to these questions, a brief account of developments in linear and nonlinear urban modelling, in optimisation and spatial interaction, and in integrated forecasting will be presented in the next section. Then the main focus of this review will be established: a general linear framework will be presented and various models such as the Lowry model will be derived as special cases. The use of linear analysis in detecting the significance of causal structure in such models, and in operational issues such as the dynamics of model calibration and solution will be noted. The emphasis will then switch to nonlinear analysis of the same models through optimisation. Through these techniques, it will be clear how similar models can be explored in quite different ways, each way enriching the other and opening up further significant research questions.

## THE DEVELOPMENT OF GENERAL URBAN MODELS.

Urban modelling has always been characterised by the development of partial models dealing with well-defined subsystems of the urban system and the use of these models as building blocks in the construction of more general models. Such partial models have been developed in depth and thus their coupling together to form more general structures has proceeded along fairly *ad hoc* lines. For example, regional economic models such as economic base and input-output have been coupled to spatial interaction models to form Lowry-type models and such coupling has only been explored in terms of the resulting more general model in the very recent past. Indeed in the development of even more comprehensive models where demographic models are linked to regional economic and spatial interaction, the integration is

even weaker and as yet there are few if any approaches enabling such structures to be developed in a direct and consistent way.

In the development of general models in this *ad hoc* way, different types of model techniques have been freely mixed. For example, the original Lowry model was stated by Lowry (1964) as an implicitly nonlinear system only to be immediately put into a linear framework with similarities to input-output analysis by Garin (1966) and Harris (1966). Wilson's (1974) development ofthe spatial interaction components was based on nonlinear optimisation through entropy-maximising but this was achieved within the linear structure of the original model which he regarded as forming the constraints and accounts characterising the framework within which spatial interaction took place. It was not until Coelho and Williams (1978) developed the model in a comprehensive nonlinear programming framework that anything with the power of the Garin linear form was established in nonlinear terms. Since then the idea of optimisation has been used to relate the Lowry model to the TOPAZ model (Sharpe and Karlquist, 1980) and to elaborate this model into various multi-activity versions (Leonardi, 1981). This theme of optimisation as a unifying and integrating feature in model design will be elaborated here in the context of the linear framework in which models such as the Lowry model have evolved. But before this review develops this idea in formal terms, a brief history of the key developments in optimisation, linear analysis and integrated forecasting will enable the context to be set.

The general form of spatial interaction model which was stated in analogy to gravitational force in physics was first developed in the formal framework ofoptimisation in the 1960's. Wilson (1967) amongst others used entropy-maximising to develop a most probable form for the model while Murchland (1966) set the model in a general mathematical programming framework which emphasised links to the more substantive question of what was

being optimised. The link between linear programming transport models and gravity models was established by Evans (1973) and used by Wilson and Senior (1974) to establish a more general nonlinear programming framework for entropy-maximising. Since then Erlander (1980) and Leonardi (1978a) have elaborated the framework in diverse ways and there have been attempts in transport to link distribution and assignment models in this way. Wilson (1982) himself has been concerned with embedding this approach in a wider dynamic context in which the activity variables between which interaction takes place, vary systematically through time.

In terms of general urban models, entropy-maximising has been used to establish consistent submodels in terms of the way locational attractions, constraints and parameters are handled but until the work of Coelho and Williams (1978), entropy-maximising was not used to generate general model structures directly. Coelho and Williams showed how Lowry-like models could be generated in a nonlinear programming framework which enabled several developments: joint estimation and solution in contrast to previous practice where estimation and solution were achieved separately, the specification of primals and their duals which enabled more efficient solution and new substantive interpretations of model parameters, and the establishment of relationships between these models and more disaggregate behavioural forms. The Coelho-Williams framework has since been built upon by Bertuglia and Leonardi (1980a) in several works and by Brotchie's group (Lesse, Brotchie, Roy and Sharpe, 1978). One interesting and somewhat eccentric early development of general models in the same spirit was Broadbent's (1973)representation of the Lowry model in an activity-commodity framework. This was never taken further but has been used recently in the design of integrated forecasting models.

The original linear framework for the Lowry model stated by Harris (1966) and Garin (1966) was suggestive in its relationship to input-output analysis. However, the model proved difficult to generalise to input-output form due to the separability of economic base relations and spatial interaction, and the general problem of reconciling economic base theory with input-output (see Romanoff, 1974). It was not until 1977 that the problem was finally resolved by Macgill(1977) who developed a demand-driven version of the Lowry model in input-output format which contrasted strongly with the conventional model in which the economic base mechanism was based on supply-driven considerations. At present, Macgill's treatment is the only formal statement of the Lowry model as an input-output model although Williams (1979) has developed an algorithmic supply-driven framework of which the Lowry model is a special case. Leonardi (1978b) and Bertuglia and Leonardi (1980a) have presented a formal version of the supply-driven framework which will be used as a starting point here. Finally, the author has used the linear framework as a vehicle to enable efficient calibration of the submodels and as a means of assessing the spatial effect of inputs on outputs, following Schinnar's work (Schinnar, 1978). These developments will be presented in later chapters in considerable detail as they form the themes of this thesis, although it is important to note them in context now.

Attempts to integrate the general model with other sectors of the urban and regional system have been achieved in more *ad hoc* ways. Gordon and Ledent (1980) show how the Lowry model can be nested within a regional framework where input-output and demographic models are integrated (see also Gordon and Ledent, 1981) and Batey and Madden (1981) develop such integration between input-output and demographic sectors using an activity-commodity framework. They also report that the Lowry model can be elaborated within such a framework. Attempts have also been made to integrate such urban

models with the conventional transport model (Echenique, 1977; Hutchinson, 1976) while extensions to incorporate household dynamics and transport behaviour at a highly disaggregate level have been pursued by Mackett (1981).

Several dynamic versions of this form of general model exist but in essence these are either 'dynamicised' versions of the static model (Said and Hutchinson, 1980; Mackett, 1981) or versions in which the model's internal mechanisms are considered as dynamic mechanisms in the temporal sense (Batty, 1976; Webber, 1979). These models all tend to be designed with operationality in mind. More recently however, Wilson (1982) and Allen, Sanglier, Boon, Deneuborg and DePalma (1981) have taken a different approach to dynamics in which activities are assumed to change in relatively tractable ways, and the system behaviour which may be surprising, is characterised by the interaction of these activities. This approach to dynamics is also offered by Wilson (1982) as a solution to the problem of matching demand and supply. In a static context, demand-supply adjustments in urban models are usually accomplished in *ad hoc* ways although Echenique, Feo, Herrera and Riquezes (1974) have developed some consistency in the use of such techniques. The development of Lowry's (1964) model using the ideas of Forrester (1969) has found little favour but there are some noteworthy applications (Burdekin, 1979; Bertuglia, Occelli, Rabino and Tadei, 1980). Of more operational interest is the recent model of Varaprasad (1980) which incorporates some of the ideas of nonlinear dynamics being pursued by Wilson.

This review will now turn to a more formal exposition of certain of the key themes alluded to above. We will first state a multi-activity linear framework which we refer to as a spatial input-output structure. Conventional models are special cases of this framework but the value of the general

model is to point to model structures which have not yet been developed but might seem relevant and feasible. Moreover such frameworks can be used to aid empirical development of appropriate model structures and to anticipate a major conclusion of this review, such theoretical-technical developments would appear essential in guiding empirical applications. Nonlinear derivations of urban models will then be reviewed and linked to the linear framework, emphasising yet another conclusion of this review - the use of linear and nonlinear frameworks simultaneously to aid model design and application. Most of these developments have occurred since the mid-1970's and the time would now seem ripe for the use of many of these ideas in an empirical context. This then will form the major conclusion to this chapter and will set the tone for the rest of this thesis.

SPATIAL INPUT-OUTPUT STRUCTURES.

We will develop a general spatial model in which activities are linked to one another in causal terms at a macro-level and linked in spatial terms at a lower level. Without loss of generality we will assume that there are N activities and I zones or spatial units over which this interaction takes place. If interactions are absent, the model can handle these without reducing the number of activities or zones (to M or J respectively). This assumption simplifies the presentation. Activities are indexed by the superscripts $n,m = 1,2,..,N$ and in locational terms by the subscript indices $i,j = 1,2,..,I$. Relationships between activities at the causal level are indexed by mn and at the spatial level by ij. The first index in each pair represents the origin or source of the relationship, the second the destination or sink. We will also assume that there are N exogenous activities in I zones and the framework is able to generate the

same number of endogenous activities.

With these assumptions, define $y_j^n$ and $x_j^n$ as the respective amounts of endogenous and exogenous activity n in zone j. These are elements in the 1 x I row vectors $\underline{y}^n$ and $\underline{x}^n$. The dependence of activity n on m is given by the scalar $\alpha^{mn}$ at the causal level while at the spatial level, the dependence is given by $A_{ij}^{mn}$ which is the typical element in the I x I matrix $\underline{A}^{mn}$. The model is based on the following general relationship:

$$y_j^n = \sum_m \alpha^{mn} \sum_i y_i^m A_{ij}^{mn} + x_j^n \quad , \quad \forall n \quad , \tag{2.1}$$

which clearly represents a set of linear simultaneous equations, soluble for $[\underline{y}^n]$, given the usual conditions on the form of the matrix $[\underline{A}^{mn}]$. Equation (2.1) displays a major property of the system, that of the separability of causal from spatial dependence. This condition can be relaxed but in the development of such models to date in an urban context, it rarely has been. Note that equation (2.1) has the same form as the framework presented by Bertuglia and Leonardi (1980a).

In this context where the central interest is on spatial distribution it is convenient to represent $\underline{y}^n$ in distributional form. Then

$$\sum_j y_j^n = 1, \ \forall n \quad . \tag{2.2}$$

The causal relations $\alpha^{mn}$ also satisfy the following conditions

$$0 \leq \alpha^{mn} \leq 1 \text{ and } 0 \leq \sum_m \alpha^{mn} \leq 1 \ . \tag{2.3}$$

As the model in equation (2.1) is concerned with transforming distributions into one another, the matrix $\underline{A}^{mn}$ is defined as a transition probability matrix or row stochastic matrix where

$$\sum_j A_{ij}^{mn} = 1 \ , \ \forall mn \tag{2.4}$$

Finally to ensure that the system is closed, the exogenous input $x_j^n$ is defined as

$$x_j^n = (1 - \sum_m \alpha^{mn}) \hat{x}_j^n \ , \tag{2.5}$$

the unweighted input $\hat{x}_j^n$ also being represented in distributional form so that

$$\sum_j \hat{x}_j^n = 1, \ \forall n. \tag{2.6}$$

Using equations (2.2) to (2.6) in (2.1), the model can now be written as

$$y_j^n = \sum_m \alpha^{mn} \sum_i y_i^m A_{ij}^{mn} + (1 - \sum_m \alpha^{mn}) \hat{x}_j^n \ , \ \forall n \tag{2.7}$$

which can easily be visualised as a flow structure. The separability of causal from spatial dependence can also be easily assessed by summing equation (2.7) over j and using the definitions in (2.2) to (2.6). If it is required to convert the exogenous or endogenous distributions into total activity form, it is only necessary to multiply these values by total activity values $\hat{X}^n$ and $\hat{Y}^n$ respectively. In this way, the outputs from this model can be linked directly to more conventional model forms such as input-output models or Lowry models.

We can represent equation (2.7) in block matrix form as

$$\underline{y}^n = \sum_m \alpha^{mn} \underline{y}^m \underline{A}^{mn} + \underline{x}^n \ , \tag{2.8}$$

or in supermatrix form as

$$\underline{y} = \underline{y} \ \underline{\Lambda} + \underline{x} \ , \tag{2.9}$$

where $\underline{y}$ is a 1 x NI row vector composed of the vectors $\underline{y}^1, \underline{y}^2, \ldots, \underline{y}^N$, $\underline{x}$ is a row vector of similar structure and dimension composed of the

vectors $\underline{x}^n$, and $\underline{\Lambda}$ is an NI x NI supermatrix, each block being formed from $\alpha^{mn}\ \underline{A}^{mn}$, $\forall mn$. Equation (2.9) is of the conventional input-output form and given the definitions in equations (2.2) and (2.6), and assuming linear independence, the solution is of the form

$$\underline{y} = \underline{x}\ (\underline{I} - \underline{\Lambda})^{-1} \tag{2.10}$$

$(\underline{I} - \underline{\Lambda})^{-1}$ is a Leontief inverse and has the usual properties of such a multiplier, that is, it can be expanded into a converging matrix series which results if equation (2.9) were to be solved with the initial starting vector for $\underline{y}$ as $\underline{x}$. Note that $\underline{I}$ is an identity matrix of appropriate order and that henceforth wherever $\underline{I}$ appears it will be such an appropriately dimensioned matrix.

Because of the separability of $\alpha^{mn}$ and $\underline{A}^{mn}$, it is possible to compute a causal multiplier which has the same structure as $(\underline{I} - \underline{\Lambda})^{-1}$. Then if equation (2.8) is aggregated by postmultiplication using the I x 1 transposed unit vector $\underline{1}^T$, then the result

$$1^n = \sum_m \alpha^{mn} + (1 - \sum_m \alpha^{mn})$$

can also be written in matrix form, the solution of which yields the causal multiplier $(\underline{I} - \underline{\alpha})^{-1}$ where $\underline{\alpha} = [\alpha^{mn}]$. In input-output analysis, the vector $\underline{x}$ which drives the system is final demand while $\underline{y}$ is a vector of production levels associated with the industries required to satisfy this demand. As mentioned above, Macgill (1977) has used this framework to elaborate a demand-driven Lowry type model where basic population provides the input. Bertuglia and Leonardi (1980a) on the other hand use the same framework to structure a supply-driven model in which basic employment is the input. Williams (1979) has provided an algorithm for solving equation (2.9) at the causal and spatial levels, while Gordon and Ledent (1980) have approximated the supermatrix form in (2.9) at separate causal (regional)

and spatial (urban) levels. As the conventional urban (Lowry) model is supply rather than demand-driven, we will concentrate on this version in the next section although Macgill's (1977) version will have the same structure as the model which we will derive.

THE LOWRY MODEL AS AN INPUT-OUTPUT STRUCTURE.

In terms of the general framework given in equations (2.8) and (2.9), the Lowry model is a two-activity spatial model, the activities being population and employment which are endogenous. The exogenous activity is basic employment and there is no exogenous population input. Before the model is elaborated, it is worth looking at the general structure of the 2 x 2 supermatrix $\underline{\Lambda}$ and its inverse $(\underline{I} - \underline{\Lambda})^{-1}$ for important simplifications can be made. First consider equation (2.10). If this is post-multiplied by the matrix $(\underline{I} - \underline{\Lambda})$ it is clear that the inverse can be written as

$$\underline{\Omega}(\underline{I} - \underline{\Lambda}) = \underline{I} \quad , \tag{2.11}$$

where $\underline{\Omega} = (\underline{I} - \underline{\Lambda})^{-1}$. Then for a 2 x 2 activity system, the explicit partitioned form of equation (2.11) becomes

$$\begin{bmatrix} \underline{\Omega}^{11} & \underline{\Omega}^{12} \\ \underline{\Omega}^{21} & \underline{\Omega}^{22} \end{bmatrix} \begin{bmatrix} \underline{I} - \alpha^{11}\underline{A}^{11} & -\alpha^{12}\underline{A}^{12} \\ -\alpha^{21}\underline{A}^{21} & \underline{I} - \alpha^{22}\underline{A}^{22} \end{bmatrix} = \begin{bmatrix} \underline{I} & \underline{0} \\ \underline{0} & \underline{I} \end{bmatrix} \tag{2.12}$$

Writing out the matrix by matrix multiplications implicit in equation (2.12) we get

$$\left.\begin{array}{lll} \underline{\Omega}^{11}(\underline{I} - \alpha^{11}\underline{A}^{11}) & - \underline{\Omega}^{12}\alpha^{21}\underline{A}^{21} & = \underline{I} \\ -\underline{\Omega}^{11}\alpha^{12}\underline{A}^{12} & + \underline{\Omega}^{12}(\underline{I} - \alpha^{22}\underline{A}^{22}) & = \underline{0} \\ \underline{\Omega}^{21}(\underline{I}-\alpha^{11}\underline{A}^{11}) & - \underline{\Omega}^{22}\alpha^{21}\underline{A}^{21} & = \underline{0} \\ -\underline{\Omega}^{21}\alpha^{12}\underline{A}^{12} & + \underline{\Omega}^{22}(\underline{I} - \alpha^{22}\underline{A}^{22}) & = \underline{I} \end{array}\right\} . \tag{2.13}$$

The equations in (2.13) can be solved in several ways for a typical element $\underline{\Omega}^{mn}$ of the inverse matrix $\underline{\Omega}$. However we can solve for each $\underline{\Omega}^{mn}$ solely in terms of the known elements $\alpha^{mn}\underline{A}^{mn}$, and for each $\underline{\Omega}^{mn}$, this is equivalent to considering the relationship mn to be central to the way the solution is generated hence interpreted. We thus refer to the following solutions for $\underline{\Omega}^{mn}$ to be the pure solution form for the inverse $\underline{\Omega}$. Its significance will be clear later. Then from equations (2.13)

$$\left.\begin{aligned}
\underline{\Omega}^{11} &= [\underline{I} - \alpha^{11}\underline{A}^{11} - \alpha^{12}\underline{A}^{12}(\underline{I} - \alpha^{22}\underline{A}^{22})^{-1}\alpha^{21}\underline{A}^{21}]^{-1} \\
\underline{\Omega}^{12} &= [(\underline{I} - \alpha^{22}\underline{A}^{22})(\alpha^{12}\underline{A}^{12})^{-1}(\underline{I}-\alpha^{11}\underline{A}^{11})-\alpha^{21}\underline{A}^{21}]^{-1} \\
\underline{\Omega}^{21} &= [(\underline{I} - \alpha^{11}\underline{A}^{11})(\alpha^{21}\underline{A}^{21})^{-1}(\underline{I} - \alpha^{22}\underline{A}^{22}) -\alpha^{12}\underline{A}^{12}]^{-1} \\
\underline{\Omega}^{22} &= [\underline{I} - \alpha^{22}\underline{A}^{22} - \alpha^{21}\underline{A}^{21}(\underline{I}-\alpha^{11}\underline{A}^{11})^{-1}\alpha^{12}\underline{A}^{12}]^{-1}
\end{aligned}\right\} . \quad (2.14)$$

We can express each matrix $\underline{\Omega}^{mn}$ in terms of any other but in the sequel, the most appropriate ways of looking at tbe solutions will be in terms of the key relationship concerning the first activity based on $\underline{\Omega}^{11}$ and the key relationship concerning the second involving $\underline{\Omega}^{22}$. We will thus write the inverse $\underline{\Omega}$ in two ways: first based on $\underline{\Omega}^{11}$ which we will call $\underline{\Omega}^{1}$ and second based on $\underline{\Omega}^{22}$ which we will call $\underline{\Omega}^{2}$. Then

$$\underline{\Omega}^{1} = \left[\begin{array}{c|c}
\underline{\Omega}^{11} & \underline{\Omega}^{11}\alpha^{12}\underline{A}^{12}(\underline{I}-\alpha^{22}\underline{A}^{22})^{-1} \\
\hline
(\underline{I}-\alpha^{22}\underline{A}^{22})^{-1}\alpha^{21}\underline{A}^{21}\underline{\Omega}^{11} & (\underline{I}-\alpha^{22}\underline{A}^{22})^{-1}[\underline{I}+\alpha^{21}\underline{A}^{21}\underline{\Omega}^{11}\alpha^{12}\underline{A}^{12}(\underline{I}-\alpha^{22}\underline{A}^{22})^{-1}]
\end{array}\right],$$

$$\underline{\Omega}^{2} = \left[\begin{array}{c|c}
(\underline{I}-\alpha^{11}\underline{A}^{11})^{-1}[\underline{I}+\alpha^{12}\underline{A}^{12}\underline{\Omega}^{22}\alpha^{21}\underline{A}^{21}(\underline{I}-\alpha^{11}\underline{A}^{11})^{-1}] & (\underline{I}-\alpha^{11}\underline{A}^{11})^{-1}\alpha^{12}\underline{A}^{12}\underline{\Omega}^{22} \\
\hline
\underline{\Omega}^{22}\alpha^{21}\underline{A}^{21}(I-\alpha^{11}\underline{A}^{11})^{-1} & \underline{\Omega}^{22}
\end{array}\right].$$

(2.15) and (2.16)

Equations (2.15) and (2.16) will be used in the following sections to display important insights into the solutions of 2 x 2 spatial activity models.

The two activities in the original Lowry (1964) model were population and

employment, the linkages between these activities being conceived in terms of the rate at which population generates service employment, and the dependence (activity rate) of population on employment. In locational terms the population is related to employment through the journey-to-work and service employment to population through the spatial demand for services. No dependence of the population on itself or employment on itself is assumed and only one activity - employment - has an exogenous component referred to as basic employment. Given these assumptions, the above framework can be written in more familiar terms as

$$\underline{y}^1 = \underline{p}\ ,\ \underline{y}^2 = \underline{e}\ ,\ \ \underline{x}^1 = \underline{0} \text{ and } \underline{x}^2 = (1-\beta)\underline{b}\ ,$$

where $\underline{p}$, $\underline{e}$ and $\underline{b}$ are 1 x I row vectors of population, total and basic employment respectively, each measured in distributional terms. $\beta$ is the ratio of service to total employment. As implied above, there is no self-dependence in the system, thus

$$\alpha^{12} = \beta\ ,\ \underline{A}^{12} = \underline{B}\ ,\quad \alpha^{21} = 1\ ,\quad \text{and } \underline{A}^{21} = \underline{C}\ ,$$

where $\underline{B}$ is the matrix of transition probabilities between population and service centres, reflecting the demand for services and $\underline{C}$ is a transition probability matrix between work and home reflecting the journey to work. The model can now be written in the form of equation (2.9) $\underline{y} = \underline{y}\,\underline{\Lambda} + \underline{x}$ which in partitioned form is

$$[\underline{p}\ \underline{e}] = [\underline{p}\ \underline{e}] \begin{bmatrix} \underline{0} & \beta\underline{B} \\ \underline{C} & \underline{0} \end{bmatrix} + [\underline{0}\ \ (1-\beta)\underline{b}]. \tag{2.17}$$

Equation (2.17) provides a very clear picture of the model's structure. Macgill's (1977) version has the same structure but the input she assumes is basic population, not basic employment, thus reflecting final demand as in input-output analysis.

The solutions to the model can now be found directly by using the definitions

prior to equation (2.17) in equations (2.14) to (2.16). First for the solution centred on treating population as the central driving force of the model, from equations (2.14) and (2.15)

$$\underline{\Omega}^{11} = [\underline{I} - \beta\, \underline{B}\, \underline{C}]^{-1} \quad , \qquad \text{and}$$

$$\underline{\Omega}^{1} = \begin{bmatrix} [\underline{I} - \beta\, \underline{B}\, \underline{C}]^{-1} & [\underline{I} - \beta\, \underline{B}\, \underline{C}]^{-1}\, \beta\, \underline{B} \\ \underline{C}[\underline{I} - \beta\, \underline{B}\, \underline{C}]^{-1} & \underline{I} + \underline{C}[\underline{I} - \beta\, \underline{B}\, \underline{C}]^{-1}\, \beta\underline{B} \end{bmatrix} .$$

From equation (2.10) in the form $\underline{y} = \underline{x}\, \underline{\Omega}^{1}$, the solutions for $\underline{p}$ and $\underline{e}$ are immediately derived

$$\underline{p} = (1-\beta)\, \underline{b}\, \underline{C}\, [\underline{I} - \beta\, \underline{B}\, \underline{C}]^{-1} \quad , \qquad \text{and} \tag{2.18}$$

$$\underline{e} = (1-\beta)\, \underline{b} + \beta(1-\beta)\, \underline{b}\, \underline{C}\, [I - \beta\underline{B}\, \underline{C}]^{-1}\, \underline{B} \quad ,$$

$$= (1-\beta)\underline{b} + \beta\, \underline{p}\, \underline{B} \quad . \tag{2.19}$$

Equations (2.18) and (2.19) are the conventional ones for the Lowry model as developed by Harris (1966), Garin (1966) and others but these reflect basic population $\underline{b}\, \underline{C}$ as the driving force. This solution appears equivalent to Macgill's model (which is stated in conventional form in Wilson, Coelho, Macgill and Williams, 1981, pp.248-249). If this casual observation is borne out by more considered reflection, this means the framework introduced here and its partitioning in the manner shown, represents a unified way of linking different types of input-output model to Lowry-like models.

The more usual form of partitioned solution is derived using equations (2.14) and 2.16). Then

$$\underline{\Omega}^{22} = [\underline{I} - \beta\underline{C}\, \underline{B}]^{-1} \quad , \qquad \text{and}$$

$$\underline{\Omega}^{2} = \begin{bmatrix} \underline{I} + \beta\, \underline{B}[\underline{I} - \beta\underline{C}\, \underline{B}]^{-1}\underline{C} & \beta\underline{B}[\underline{I} - \beta\underline{C}\, \underline{B}]^{-1} \\ [\underline{I} - \beta\, \underline{C}\, \underline{B}]^{-1}\underline{C} & [\underline{I} - \beta\underline{C}\, \underline{B}]^{-1} \end{bmatrix} .$$

from which the solutions for $\underline{p}$ and $\underline{e}$ using $\underline{y}$ and $\underline{x}\ \underline{\Omega}^2$ are

$$\underline{p} = (1-\beta)\underline{b}\ [\underline{I} - \beta\ \underline{C}\ \underline{B}]^{-1}\underline{C} \quad , \qquad \text{and} \qquad (2.20)$$

$$\underline{e} = (1-\beta)\underline{b}\ [I - \beta\ \underline{C}\ \underline{B}]^{-1} \quad . \qquad (2.21)$$

Equations (2.20) and (2.21) are those originally stated by Garin (1966) reflecting the basic employment driven model. The inverses in equations (2.18) to (2.21) can both be expanded as Leontief series which are indicative of the generation of activity in sequential form; in the case of the multiplier $\underline{\Omega}^{11}$ starting from basic population $(1-\beta)\underline{b}\ \underline{C}$ and in the case of $\underline{\Omega}^{22}$ starting from basic employment $(1-\beta)\underline{b}$. In another sense, equations (2.20) and (2.21) can be seen as 'duals' of equations (2.18) and (2.19) respectively. This interpretation is aided by noting that the transformations in this model are only of population into employment and vice versa.

GENERALISED LOWRY MODELS.

In the conventional model, there is only one exogenous input - basic employment, and no feedback (dependence) within each of the two sectors. Both these assumptions can be relaxed in generalising the model. First assume that there is an exogenous distribution of population - basic population $\underline{h}$ which is incorporated into the input vector $\underline{x}^1$ as

$$\underline{x}^1 = (1-\gamma)\underline{h} \text{ and } \underline{x} = [(1-\gamma)\underline{h},\ (1-\beta)\underline{b}] \quad .$$

$\gamma$ is the ratio of non-basic or endogenous to total population and is set to $\alpha^{21}$. The matrix $\underline{\Lambda}$ thus becomes

$$\underline{\Lambda} = \begin{bmatrix} \underline{0} & \beta\underline{B} \\ \gamma\underline{C} & \underline{0} \end{bmatrix}$$

For the inverse form based on $\underline{\Omega}^1$, the population and employment equations are

$$\underline{p} = [(1-\gamma)\underline{h} + \gamma(1-\beta)\underline{b}\ \underline{C}][\underline{I} - \gamma\beta\underline{B}\ \underline{C}]^{-1} \quad , \qquad \text{and} \tag{2.22}$$

$$\underline{e} = \beta[(1-\gamma)\ \underline{h} + \gamma(1-\beta)\underline{b}\ \underline{C}][\underline{I} - \gamma\beta\underline{B}\ \underline{C}]^{-1}\underline{B} + (1-\beta)\underline{b} \quad . \tag{2.23}$$

These population-driven solutions can be contrasted with their duals derived using $\underline{\Omega}^2$ which are stated as

$$\underline{p} = (1-\gamma)\underline{h} + [\beta(1-\gamma)\underline{h}\ \underline{B} + (1-\beta)\underline{b}][\underline{I} - \gamma\beta\underline{C}\ \underline{B}]^{-1}, \text{ and} \tag{2.24}$$

$$\underline{e} = [\beta(1-\gamma)\underline{h}\ \underline{B} + (1-\beta)\underline{b}][\underline{I} - \gamma\beta\ \underline{C}\ \underline{B}]^{-1} \quad . \tag{2.25}$$

The two inverses in equations (2.22), (2.23) and (2.24), (2.25) can be expanded in the usual way reflecting the effects of inputs of population and employment from exogenous sources respectively. There are a number of ways of looking at these duals. Clearly the ratios $\gamma$ and $\beta$ reflect the importance (size) of input populations and employments and it is clear that the other extreme to the conventional basic employment-driven Lowry model - the basic population-driven model - is derived when $\beta = 1$. In such a case, equations (2.22) and (2.23) are the most appropriate forms. It is surprising that no applications (to the author's knowledge, that is) of models in which basic population features, have been developed, for there would appear to be many situations where this might apply. Indeed the great value of the framework presented earlier is its ability to enable generalisation of model structures and to point out 'obvious' model types which have hitherto been disregarded. Only of late have such generalisations been attempted and although used for example by Bertuglia and Leonardi (1980a), the properties of the framework have not been made explicit. In the penultimate chapter of this thesis, this general framework will be elaborated empirically when spatial invariance and generalised model structures are examined.

We will now generalise the model further and look at the case where self-dependence of activities is assumed. Then noting that

$$\alpha^{11} = \alpha, \quad \underline{A}^{11} = \underline{A}, \quad \alpha^{22} = \sigma, \text{ and } \underline{A}^{22} = \underline{D},$$

the input vectors $\underline{x}^1$ and $\underline{x}^2$ become

$$\underline{x}^1 = (1-\alpha-\gamma)\underline{h} \text{ and } \underline{x}^2 = (1-\beta-\sigma)\,\underline{b} \quad .$$

The model is now

$$[\underline{p}\ \underline{e}] = [p\ e] \begin{bmatrix} \alpha\underline{A} & \beta\underline{B} \\ \gamma\underline{C} & \sigma\underline{D} \end{bmatrix} + [(1-\alpha-\gamma)\underline{h}\ (1-\beta-\sigma)\underline{b}] \ , \qquad (2.26)$$

where equation (2.26) has a complete structure. We will examine the solution to this model using the inverse $\underline{\Omega}^2$ based on $\underline{\Omega}^{22}$ although this is now arbitrary for such a complete structure. Then

$$\underline{\Omega}^{22} = [I - \sigma\underline{D} - \gamma\underline{C}\ (\underline{I} - \alpha\underline{A})^{-1}\ \beta\underline{B}]^{-1} \quad , \qquad (2.27)$$

which can be expanded first as

$$\underline{\Omega}^{22} = [\underline{I} + (\sigma\underline{D} + \gamma\underline{C}\ (\underline{I} - \alpha\underline{A})^{-1}\beta\underline{B}) + (\sigma\underline{D} + \gamma\underline{C}(\underline{I} - \alpha\underline{A})^{-1}\beta\underline{B})^2 + \ldots.]. \qquad (2.28)$$

Clearly equation (2.28) can be further expanded in terms of $(\underline{I} - \alpha\underline{A})^{-1}$ and this shows the confounded nature of the generation process. For example, an input of employment generates self-employment through $\sigma\underline{D}$, then population through $\gamma\underline{C}$ which in turn generates a whole series of self-populations leading to new employment, further self-induced employment and so on. The process no longer has the simplicity of the expansions associated with the Lowry model but it does emphasise the need to think deeply about the causal relations associated with a structure as simple as this one based on only two activities. Moreover, it highlights the need for some means of assessing the importance of such causal relationships in any empirical application. For completeness, we will state the solutions based on $\underline{\Omega}^{22}$ for population

and employment. Then

$$\underline{p} = (1-\alpha-\gamma)\underline{h}\ (\underline{I}-\alpha\underline{A})^{-1} + [(1-\alpha-\gamma)\underline{h}(\underline{I}-\alpha\underline{A})^{-1}\beta\underline{B} + (1-\beta-\sigma)\underline{b}]\underline{\Omega}^{22}\gamma\underline{C}(I-\alpha\underline{A})^{-1}, \quad (2.29)$$

$$\underline{e} = [(1-\alpha-\gamma)\underline{h}(\underline{I} - \alpha\underline{A})^{-1}\ \beta\underline{B} + (1-\beta-\sigma)\underline{b}]\ \underline{\Omega}^{22} \quad . \quad (2.30)$$

Equations (2.29) and (2.30) display the complexity of effects which increase exponentially with the number of distinct activities comprising the model.

The complete model structure based on two activities is hard to test empirically due to the difficulties inherent in unravelling feedback effects and clearly the full model with N activities would present a major estimation problem. As far as the author knows, a model with more than two activities considered in the framework developed here, has not been applied empirically although there have been hybrid versions such as that due to Geraldes, Echenique and Williams (1978) which make use of Williams' (1979) algorithm for spatial input-output analysis. A more favoured strategy for elaborating these types of model structure has been through disaggregation (see Wilson, 1974) but this has not generally led to new and different structures.

The major conclusion from the generalisations introduced so far is the need for a high degree of discrimination concerning appropriate model structures. Indeed, the success of the original Lowry model may well be largely due to its parsimonious representation of causal relationships and its emphasis on the most significant ones through *ad hoc* developments. One obvious simplification of the complete two activity model which has been suggested relates to the self-dependence effects. The matrices $\underline{A}$ and $\underline{D}$ could be considered as identity matrices, that is,

$$\underline{A} = \underline{I} \text{ and } \underline{D} = \underline{I} \quad ,$$

if it is assumed that the population and employment dependences do not have spatial effects. This seems logical for population where dependence might relate to the generation of households but less so for services, unless the zoning system is chosen to pick up just those services which locate adjacent to those generated from the demands of the population.

Noting then that $(\underline{I} - \alpha\underline{A})^{-1} = (1-\alpha)^{-1}\,\underline{I}$ and $\sigma\underline{D} = \sigma\underline{I}$, with these simplifications, the multiplier in equation (2.27) becomes

$$\underline{\Omega}^{22} = [\underline{I} - (\sigma\underline{I} + (1-\alpha)^{-1}\gamma\beta\ \underline{C}\ \underline{B})]^{-1} \quad . \tag{2.31}$$

Only the employment equation need be stated here and this, using equation (2.31) is derived as

$$\underline{e} = [\beta(1-\alpha)^{-1}(1-\alpha-\gamma)\underline{h}\ \underline{B} + (1-\beta-\gamma)\underline{b}]\underline{\Omega}^{22} \tag{2.32}$$

Equation (2.32) thus has a very similar structure to equation (2.25) which allows meaningful elaboration in series form and has smaller data requirements. Moreover the simplifications introduced here also show once again the importance of the separability of economic base relations from spatial interaction in this framework.

To complete this section a simple example of how the framework might be used to design different model structures which exhibit the property of parsimony, is worth illustrating. One set of structures which are tractable is given by a closed chain of activities with no cross or self-dependences. Using this idea, consider a model driven by basic employment in which the distribution of demand for services from the population, given by the 1 x I row vector $\underline{e}^1$ is different from its supply $\underline{e}^2$. The demand is generated from the population as $\underline{e}^1 = \underline{p}\ \underline{B}$ and there is a spatial interaction matrix $\underline{D}$ which converts a fraction $\sigma$ of this demand into supply and adds it to the given fraction of basic employment $(1-\sigma)\underline{b}$ as $\underline{e}^2 = \sigma\underline{e}^1\underline{D} + (1-\sigma)\underline{b}$. Thus $\underline{D}$

is a matrix linking demand to supply which detects disequilibrium effects in the system. The chain is then closed in that the distribution of population is generated from the supply of employment as $\underline{p} = \underline{e}^2\underline{C}$. This three activity model can now be written as

$$[\underline{p}\ \underline{e}^1\underline{e}^2] = [\underline{p}\ \underline{e}^1\underline{e}^2] \begin{bmatrix} \underline{0} & \underline{B} & \underline{0} \\ \underline{0} & \underline{0} & \sigma\underline{D} \\ \underline{C} & \underline{0} & \underline{0} \end{bmatrix} + [\underline{0}\ \underline{0}\ (1-\sigma)\underline{b}] \quad . \qquad (2.33)$$

The solution to equation (2.33) using the pure form inverses are

$$\left.\begin{aligned} \underline{p} &= (1-\sigma)\underline{b}\ \underline{C}\ [\underline{I} - \sigma\ \underline{B}\ \underline{D}\ \underline{C}]^{-1} \\ \underline{e}^1 &= (1-\sigma)\underline{b}\ \underline{C}\ \underline{B}\ [\underline{I} - \sigma\underline{D}\ \underline{C}\ \underline{B}]^{-1}\ , \qquad \text{and} \\ \underline{e}^2 &= (1-\sigma)\underline{b}\ [\underline{I} - \sigma\underline{C}\ \underline{B}\ \underline{D}]^{-1} \end{aligned}\right\} . \qquad (2.34)$$

The matrices $\underline{B}\ \underline{D}\ \underline{C}$, $\underline{D}\ \underline{C}\ \underline{B}$ and $\underline{C}\ \underline{B}\ \underline{D}$ give the overall spatial interaction patterns between the same activities in the chain, but the matrices $\underline{B}\ \underline{D}$, $\underline{D}\ \underline{C}$ and $\underline{C}\ \underline{B}$ as well as the original matrices all represent patterns which enable calibration of the model and assessment of the significance of the causal structure adopted. Furthermore as $\underline{D}$ is a measure of disequilibrium, the model might easily be cast in a dynamic framework in which $\underline{D}$ is hypothesised to be a function of the mismatch between demand and supply, $\underline{e}^1$ and $\underline{e}^2$. If the dynamics enable an equilibrium to be reached in which $\underline{D} \rightarrow \underline{I}$, then it is easy to show that the model collapses back to the original Lowry model in equations (2.18) to (2.21) although in the structure given in equation (2.34), service employment $\underline{e}^1$ is considered separately from total employment $\underline{e}^2$. Throughout this section, the need to test the significance and relevance of the postulated structure has been paramount, and using the linear framework, it is possible to embark on useful tests in the manner shown in the next section.

EMPIRICAL IMPLICATIONS OF THE LINEAR MODEL FRAMEWORK.

The linear framework elaborated so far mainly serves to emphasise and clarify questions of causality concerning the model's structure. In general, such questions can only be resolved in terms of the model's underlying theory, and as already argued, it is not the purpose of this thesis to research such substantive questions. However, the linear framework has other implications, particularly in the use of linear analysis to aid in model estimation and solution. Exploring these ideas will be the main quest of this thesis and in this section, an indication of the empirical implications of causal model structure for the analysis of spatial variation will be given. The full power of this analysis in terms of measurement and estimation, and consequent interpretation will be addressed in the following section.

We will first examine the importance of the inputs to the two activity model in spatial distributional terms, and to this end, we will first examine the model as given above in equation (2.17). In this model, consider the case where $\beta = 1$, and thus the input $\underline{x} = [\underline{0}\ \underline{0}]$ is absent from equation (2.17). The model thus simplifies to

$$[\underline{p}\ \underline{e}] = [\underline{p}\ \underline{e}] \begin{bmatrix} \underline{0} & \underline{B} \\ \underline{C} & \underline{0} \end{bmatrix} , \qquad (2.35)$$

but the inverse form solution no longer applies. To solve equation (2.35) then, consider iteration starting from an arbitrary distribution vector $[\underline{p}\ \underline{e}]^0$. Then for iteration $2\tau$, the following recurrence relations hold

$$\left.\begin{aligned} [\underline{p}\ \underline{e}]^{2\tau} &= [\underline{p}\ \underline{e}]^0 \begin{bmatrix} (\underline{B}\ \underline{C})^{\tau} & \underline{0} \\ \underline{0} & (\underline{C}\ \underline{B})^{\tau} \end{bmatrix} , \\ [\underline{p}\ \underline{e}]^{2\tau-1} &= [\underline{p}\ \underline{e}]^0 \begin{bmatrix} \underline{0} & (\underline{B}\ \underline{C})^{\tau-1}\underline{B} \\ (\underline{C}\ \underline{B})^{\tau-1}\underline{C} & \underline{0} \end{bmatrix} . \end{aligned}\right\} \qquad (2.36)$$

If we examine $(\underline{C}\ \underline{B})^{\tau}$ and note that this matrix is stochastic and strongly-connected in the graph theoretic sense (due to our assumption of a connected spatial system), then in the limit $(\underline{C}\ \underline{B})^{\tau}$ converges to

$$\lim_{\tau \to \infty} (\underline{C}\ \underline{B})^{\tau} \to \underline{Z} \qquad ,$$

where $\underline{Z}$ is a row stochastic matrix in which each row is identical. $\underline{Z}$ is called an idempotent matrix in that upon further multiplication by $\underline{C}\ \underline{B}$ it is stable and unchanging, that is $\underline{Z} = \underline{Z}\ \underline{C}\ \underline{B}$. Using this result, equations (2.36) can be examined in their partitioned form in the limit and these become

$$\lim_{\tau \to \infty} \left.\begin{array}{llll} \underline{p}(2\tau) & = \underline{p}(0)(\underline{B}\ \underline{C})^{\tau} & = \underline{p}(0)\underline{Z}\ \underline{C} & = \tilde{\underline{p}} \\ \underline{e}(2\tau) & = \underline{e}(0)(\underline{C}\ \underline{B})^{\tau} & = \underline{e}(0)\underline{Z} & = \tilde{\underline{e}} \\ \underline{p}(2\tau-1) & = \underline{e}(0)(\underline{C}\ \underline{B})^{\tau-1}\underline{C} & = \underline{e}(0)\underline{Z}\ \underline{C} & = \tilde{\underline{p}} \text{ and} \\ \underline{e}(2\tau-1) & = \underline{p}(0)(\underline{B}\ \underline{C})^{\tau-1}\underline{B} & = \underline{p}(0)\underline{Z} & = \tilde{\underline{e}} \end{array}\right\} \quad (2.37)$$

where $\tilde{\underline{p}}$ is a row of $\underline{Z}\ \underline{C}$ and $\tilde{\underline{e}}$ is a row of $\underline{Z}$.

From equation (2.37) and the properties of the idempotent matrices $\underline{Z}\ \underline{C}$ and $\underline{Z}$, the solutions to equation (2.35) in partitioned form are

$$\tilde{\underline{p}} = \tilde{\underline{p}}\ \underline{B}\ \underline{C} \qquad \text{and} \qquad \tilde{\underline{e}} = \tilde{\underline{e}}\ \underline{C}\ \underline{B} \quad . \qquad (2.38)$$

In short, the iterations on equation (2.36) are dual Markov processes. The model without inputs is structurally equivalent to Coleman's (1973) model of collective action based on the theory of social exchange, and this has been explored in terms of the structure of its interaction matrices by the author (Batty, 1981a). Moreover, the special case derived here is equivalent to a Lowry model without inputs which as intuition suggests, predicts population and employment to be a function solely of the interaction pattern. This in itself is a model worth exploring further.

If the model's predictions in the case of no exogenous variables are entirely a function of the interaction matrices, the question must be asked as to what extent the model's predictions are a function of the interaction patterns when inputs are present. In the situation where $\underline{C}\ \underline{B}$ is already idempotent, then it is clear that the inputs would have no spatial impact on that portion of an activity which is generated endogenously. Thus to explore the question, it is necessary to see how close $\underline{C}\ \underline{B}$ is to its steady state form $\underline{Z}$ or to any other steady state pattern. To proceed, let us first consider the case where $\underline{C}\ \underline{B}$ is already in the steady state, that is where

$$\underline{C}\ \underline{B} = (\underline{C}\ \underline{B})^{\tau} = \underline{Z}, \quad \tau \geqslant 1 .$$

Then examining the traditional Lowry model in terms of its employment generation given in equation (2.21), the multiplier $[\underline{I} - \beta\underline{C}\ \underline{B}]^{-1}$ can be simplified as follows:

$$\begin{aligned}[\underline{I} - \beta\underline{C}\ \underline{B}]^{-1} &= \lim_{\tau \to \infty} [\underline{I} + \beta\underline{C}\ \underline{B} + (\beta\underline{C}\ \underline{B})^2 + \ldots + (\beta\underline{C}\ \underline{B})^{\tau}], \\ &= \underline{I} + \beta(1-\beta)^{-1}\ \underline{Z}. \end{aligned} \tag{2.39}$$

In the sequel, we will just examine the employment equation for the two activity model, for the population equation is subject to the same type of analysis. Using equation (2.39) in (2.21), the employment $\underline{e}$ is now written as $\hat{\underline{e}}$ and referred to as the steady state employment with input. Then

$$\begin{aligned}\hat{\underline{e}} &= (1-\beta)\underline{b}[\underline{I} + \beta(1-\beta)^{-1}\underline{Z}] \quad , \\ &= (1-\beta)\underline{b} + \beta\underline{b}\ \underline{Z} = (1-\beta)\underline{b} + \beta\tilde{\underline{e}} \quad . \end{aligned} \tag{2.40}$$

Equation (2.40) shows that $\hat{\underline{e}}$ is clearly composed of the exogenous input and an endogenous term $\beta\tilde{\underline{e}}$ which is the steady state employment from equation (2.38), the model with no input, which is entirely independent of basic employment in spatial terms.

In short, the input has no effect on the model and is thus irrelevant in a spatial sense. A similar result holds for the Lowry model with population and employment inputs. Using these results, equation (2.25) becomes

$$\begin{aligned} \hat{\underline{e}} &= [\beta(1-\gamma)\underline{h}\ \underline{B} + (1-\beta)\underline{b}][\underline{I} + \gamma\beta(1-\gamma\beta)^{-1}\underline{Z}], \\ &= \beta(1-\gamma)\underline{h}\ \underline{B} + (1-\beta)\underline{b} + \gamma\beta\ \tilde{\underline{e}} \quad . \end{aligned} \tag{2.41}$$

The critical issue here is how the predictions from the actual model compare with its steady state equivalents $\hat{\underline{e}}$ and $\tilde{\underline{e}}$, say. To explore this, we must now see how close $\underline{C}\ \underline{B}$ is to $\underline{Z}$ for if $\underline{C}\ \underline{B} = \underline{Z}$, the model is spatially invariant to its input and the input irrelevant. This possibility was first noted by Schinnar (1978) and a full formal analysis is developed in Chapter 10 after considerable empirical analysis of the model's solution dynamics in the earlier chapters has set the context. Then in Chapter 11, an empirical analysis for Melbourne is attempted which reveals a high degree of spatial invariance in model predictions in terms of inputs. This possibility is clearly evident in spatial systems which are highly polarised, that is dominated by city centres, say, and many previous applications are cast in doubt by these findings. Moreover these ideas can be used positively in the design of relevant zoning systems which will capture essential variation.

## LINEAR ANALYSIS OF SPATIAL VARIATION AND MODEL ESTIMATION.

To anticipate the subsequent analysis, results from stochastic matrix theory taken from Bailey (1964) and Bartholomew (1982), and which will be presented in detail again later, must now be presented. Any strongly-connected row stochastic matrix $\underline{P}$ can be expressed as an additive sum of its steady state form $\lim_{\tau \to \infty}(\underline{P})^{\tau} = \underline{Z}$ and deviations from this steady state.

The matrix can be represented as the sum of its eigenvalues and eigenvectors. the so-called spectral decomposition, as

$$\underline{P} = \sum_{j=1}^{I} \lambda_j \underline{V}_j \quad , \tag{2.42}$$

where it is assumed that there are I distinct eigenvalues $\lambda_j$ for the I x I matrix and that $\underline{V}_j$ is a matrix associated with the $\lambda_j$ eigenvalue constructed from suitably scaled right- and left-hand eigenvectors, $\underline{r}_j$ and $\underline{s}_j$, of $\underline{P}$. The matrix $\underline{V}_j$ is formed as $\underline{V}_j = \underline{r}_j^T \underline{s}_j$ and the scales of $\underline{r}_j$ and $\underline{s}_j$ are chosen so that $\underline{V}_\ell \underline{V}_j = \underline{0}, \ell \neq j$, $\underline{V}_\ell \underline{V}_j = \underline{V}_j$, $\ell = j$ and $\Sigma_{j=1}^{I} \underline{V}_j = 1$. If the eigenvalues of $\underline{P}$ are ordered so that $\lambda_1(=1) > |\lambda_2| > \ldots > |\lambda_I|$, then $\underline{P}$ in equation (2.42) has the following property

$$\underline{P}^\tau = \sum_{j=1}^{I} \lambda_j^\tau \underline{V}_j \quad . \tag{2.43}$$

As $\lambda_1 \underline{V}_1 = \underline{Z}$, that is, that the dominant eigenvalue and vectors determining the steady state matrix, equation (2.43) can be expressed as the sum of the steady state and deviations from it. Then

$$\underline{P}^\tau = \underline{Z} + \sum_{j=2}^{I} \lambda_j^\tau \underline{V}_j \quad , \qquad \text{and} \tag{2.44}$$

$$\underline{P} - \underline{Z} = \sum_{j=2}^{I} \lambda_j \underline{V}_j \quad . \tag{2.45}$$

As $\tau \to \infty$, equation (2.44) converges to $\underline{Z}$ which implies that the deviations in equation (2.45) converge to the zero matrix. We are now in a position to use these results to determine a decomposition for any matrix series in which $\underline{P}$ is a row stochastic matrix and $\beta$ a ratio between zero and one. Then

$$[\underline{I} - \beta\underline{P}]^{-1} = \sum_{\tau=0}^{\infty} \beta^\tau \underline{P}^\tau = \underline{I} + \beta(1-\beta)\underline{Z} + \sum_{j=2}^{I} \beta\lambda_j(1-\beta\lambda_j)^{-1}\underline{V}_j, \tag{2.46}$$

which is the series representation used by Bartholomew (1982) for manpower

planning models which have a similar structure. Quite clearly, equation (2.46) is composed of three effects: an input effect $\underline{I}$, a steady-state effect based on $\underline{Z}$ and a deviation effect based on $\underline{V}_j$.

These results in equations (2.42) and (2.46) can now be used to relate the original two activity Lowry model to its steady-state equivalents given in equations (2.38) and (2.40). Noting now that $\underline{P}$ in equation (2.46) is $\underline{C}\ \underline{B}$, and that $\underline{Z}, \lambda_j$ and $\underline{V}_j$ now pertain to $\underline{C}\ \underline{B}$, equation (2.21) can be written as

$$\underline{e} = (1-\beta)\underline{b}\ [\underline{I} + \beta(1-\beta)^{-1}\ \underline{Z} + \sum_{j=2}^{I} \beta\lambda_j(1-\beta\lambda_j)^{-1}\ \underline{V}_j] \quad ,$$

$$= (1-\beta)\underline{b} + \beta\underline{\tilde{e}} + (1-\beta)\underline{b} \sum_{j=2}^{I} \beta\lambda_j(1-\beta\lambda_j)^{-1}\ \underline{V}_j \quad . \qquad (2.47)$$

The last term on the RHS of the second line of (2.47) is the deviation from the steady state with input and it is this effect which measures the degree to which the input $(1-\beta)\underline{b}$ influences the final spatial distribution of employment. An aggregate picture of this distortion from the steady state is given by $\underline{e}-\underline{\hat{e}}$, and it is clear that wherever a series of the form in equation (2.46) appears in these models, the same type of analysis can be invoked. The same type of analysis can be developed for the two input model in equations (2.22) to (2.25), and at a more detailed level in terms of the original matrices $\underline{B}$ and $\underline{C}$ and this shows again the power of the linear framework in designing theoretical models with relevant empirical applications.

Generalising this analysis to the complete two activity model as specified in equations (2.26) to (2.30) is fairly straightforward although a full algebraic presentation would be fairly cumbersome. Thus only a qualitative discussion is developed here. For example, concentrating on the employment

equation (equation (2.30)), the inner inverse $[\underline{I} - \alpha\underline{A}]^{-1}$ can be decomposed in the manner of equation (2.46) and thus the population input $\underline{h}$ can first be separated into true input and steady state effects due to self-dependence effects generated within the multiplier $\underline{\Omega}^{22}$ and this enables the series based on $\underline{\Omega}^{22}$ to be decomposed. Finally the portion of this series remaining also forms a sub-series involving only $\gamma\underline{C}\beta\underline{B}$ and this in turn can be expressed in the manner of equation (2.46). Even within this analysis, several different types of approach can be taken by concentrating on the original interaction matrices or on cross effects, or by using different criterion for the measurement of spatial invariance-idempotence.

For models based on more than two activities, algebraic analysis becomes increasingly laborious as does representing solutions in partitioned matrix form. In such situations, it would appear that an algorithm is required for tracing through the effects of idempotence. Such an idea could be easily implemented in any empirical application. Alternatively idempotence could be assumed in different interaction patterns and comparisons then made between different model solutions at an aggregate level. Finally, the possibilities for developing these ideas at the higher level of the framework, in terms of the overall structure $\underline{y} = \underline{y}\,\underline{\Lambda} + \underline{x}$, seem slim as $\underline{\Lambda}$ does not have the appropriate Markov form.

The analysis of spatial invariance is first anticipated in Chapters 6 and 7 but is only formally and empirically developed in Chapters 10 and 11. Before that this thesis embarks on another application of linear analysis in which the dynamics of model solution - estimation and calibration - are explored through a dynamic elaboration of the models linear structure. Such a dynamic elaboration is not in terms of real-time dynamics although

there are hints of this in what follows, but in terms of solution dynamics which affect computer time, iteration time and so on. Various elaborations which lead to different model types and algorithms for their solution are presented starting in Chapter 3 and merging by Chapters 9 and 10 into the analysis of spatial invariance.

To give some feel for this work, consider the problem of estimating the parameters of the models governing the general model's spatial interaction patterns. If as is usual practice, the matrices $\underline{A}$, $\underline{B}$, $\underline{C}$, and $\underline{D}$ are formed from spatial interaction models, the parameters of these models need to be estimated, and this usually involves some iterative scheme (Batty, 1976). Such iterative methods are usually invoked prior to solution of the linear model framework, or the linear framework is nested within a wider nonlinear iterative calibration scheme. However, it is possible to solve the linear model iteratively, for example, by working out each term in the inverse expansions, and this in fact appears to be normal practice. The calibration method which builds on the work of the later chapters involves matching these two iterative processes; that is, using a single iterative scheme to enable model solution and calibration to be achieved simultaneously.

To demonstrate the idea, consider the original two activity Lowry model in equation (2.17) in terms of the equilibrium employment equation. Then it is clear that

$$\underline{e} = \beta \underline{e}\ \underline{C}\ \underline{B} + (1-\beta)\underline{b} \quad . \tag{2.48}$$

One way of solving equation (2.48) is to start with some estimate of $\underline{e}$ on the RHS of the equation and iterate the solution until a convergence limit has been met. This is the scheme adopted by Baxter and Williams (1975)

and Wilson, Coelho, Macgill and Williams (1981) to solve the Lowry model and it can be presented as

$$\underline{e}\ (\tau) = \beta\ \underline{e}\ (\tau-1)\ \underline{C}\ \underline{B} + (1-\beta)\ \underline{b} \quad . \tag{2.49}$$

As already mentioned $\underline{C}$ and $\underline{B}$ need to be estimated prior to the use of equation (2.49) or equation (2.49) is set within some wider process of estimating $\underline{C}$ and $\underline{B}$. An efficient strategy has been developed for estimating $\underline{C}$ and $\underline{B}$ using direct iteration on equation (2.49). Formally then

$$\underline{e}(\tau) = \beta\underline{e}(\tau-1)\underline{C}(\tau-1)\underline{B}(\tau-1)+(1-\beta)\underline{b} \quad , \tag{2.50}$$

where $\underline{C}(\tau)$ and $\underline{B}(\tau)$ are functions involving the distribution of employment $\underline{e}(\tau)$. A variety of schemes are developed in later chapters to enable efficient solution and estimation to be achieved simultaneously using this idea and one consequence of equations (2.48) and (2.50) is that if $\underline{C}(\tau)$ and $\underline{B}(\tau)$ converge to stable matrices $\underline{C}$ and $\underline{B}$ before $\underline{e}$ is attained, equation (2.50) collapses back to equation (2.49), thence (2.48). The advantages of this structure are exploited fairly intensively later and have been used in a more substantive context by Berechman (1976).

Although this has not yet been accomplished, it would seem quite straightforward to generalise these ideas to the higher level and to simultaneously solve and estimate the supermatrix equation $\underline{y} = \underline{y}\ \underline{\Lambda} + \underline{x}$ in analogous fashion. In models such as these usually the elements $\alpha^{nm}$ are assumed given but even these may be subject to estimation. This type of generalisation is a direct result of specifying Lowry-like models in the more general framework of a spatial input-output structure, and it represents the point that many ideas developed for two or even single activity models can be applied to multi-activity systems. This point is reinforced in the next section where we turn to nonlinear analysis of the same types of structure.

## NONLINEAR OPTIMISATION MODELS.

So far in this review, we have emphasised model structure largely through the embodiment of causal and spatial relations as linear accounts. We have emphasised how such structural questions should be explored empirically through relating inputs to outputs but we have not dealt with the form of the spatial relations assumed for these models. Only at the end of the last section was any hint given that such spatial relations might be modelled, although there has been a tacit assumption throughout this chapter that such interactions do embody nonlinear relations. In this section, we are going to turn the linear model framework inside out and study it from the point of view of modelling spatial interactions. As is well-known, such spatial interaction models are intrinsically nonlinear, and the nonlinear framework which results, will have different properties from that developed above. However one central result of this section will be to show that nonlinear multi-activity spatial models can be cast into the linear framework to enable the empirical power of that framework to be of use in evaluating model structures. This interchangeability of linear and nonlinear is a major achievement of the field over the last decade.

First developments of the Lowry model were briefly reviewed in an earlier section and it was largely Wilson's (1974) achievement to enable consistent spatial interaction models - residential location and service centre-shopping models - to be embedded within the overall model framework. In particular, the rigour imposed by entropy-maximising enabled the process of building and estimating consistent models subject to realistic constraints to be handled comprehensively and efficiently. These early developments however paid little regard to model structure, and it was not until the

impetus of treating entropy-maximising methods as special cases within nonlinear mathematical programming really began, that much thought was given to how comprehensive model structures could be derived as generalised mathematical programming problems.

Parallel to the concern for generalised optimisation methods was the concern over what was being optimised. The relationship between linear programming transport (Evans, 1973) and land use (Herbert and Stevens, 1960) models and entropy-maximising spatial interaction models focussed the question on the costs and benefits of interaction. Enormous strides have been made in relating objective functions such as entropy to consumer surplus, diversity, dispersion, variety, utility, accessibility and related measures of welfare, and the analysis of mathematical programming duals has enabled a clear picture of the cost-benefit structure of these models to be established (Harris, 1979; Williams and Senior, 1978). The Leeds group under Wilson (Wilson, Coelho, Macgill and Williams, 1981), the CSIRO group under Brotchie (Lesse, Brotchie, Roy and Sharpe, 1978; Brotchie and Lesse, 1979) and the Turin group under Bertuglia and Leonardi (1979) have done much to develop these ideas. We will not review these exciting developments here but they remain an integral part of the more technical issues emphasised in this chapter.

Coelho and Williams (1978) were the first to present a nonlinear programming derivation of the Lowry model although subsequently Leonardi (1978a) and Sharpe and Karlquist (1980) have developed similar versions. Here we will present the Coelho-Williams model and in the next section generalise to the complete two activity model. We will also show how these models can be easily cast back into the linear framework, thus enabling the previous analyses to be applied. A couple of notational details must be clarified. Here we will keep to the strict rule that the first subscript

of any variable relates to the origin of the variable in spatial terms, the second subscript to the destination. All interaction variables will be origin-constrained. We will not follow the Coelho-Williams notation in which the subscript i is reserved for population zones and j for employment zones for reasons which will be obvious when we generalise their model.

In the original Lowry Model, two interaction variables $t_{ij}$, the probability of working in i and living in j, and $s_{ij}$, the probability of living in i and demanding services in j, are required. These variables are subject to the following origin constraints

$$\sum_j t_{ij} = e_i \quad , \qquad \text{and} \tag{2.51}$$

$$\sum_j s_{ij} = p_i \quad , \tag{2.52}$$

where $e_i$ and $p_i$ are employment and population respectively defined as earlier in distributional terms so that

$$\sum_i e_i = \sum_i p_i = 1 \quad .$$

From equation (2.17), the economic base relations can be written as

$$p_j = \sum_i t_{ij} \quad , \qquad \text{and} \tag{2.53}$$

$$e_j = \beta \sum_i s_{ij} + (1-\beta) b_j \quad . \tag{2.54}$$

The model is thus subject to constraint euqations (2.51) to (2.54) which can be viewed as both origin and destination constraints on $\{t_{ij}\}$ and $\{s_{ij}\}$, thus enabling the model to be seen as two interlocking gravity models (Wilson, Coelho, Macgill and Williams, 1981).

Coelho and Williams (1978) now set up an objective function involving $\{t_{ij}\}$ and $\{s_{ij}\}$ which can be optimised with respect to these interaction variables subject to known constraints. They choose various objective functions, in particular a group surplus function which is consistent with micro-behavioural considerations, but a similar and perhaps more conventional function leading to an equivalent model is the group entropy function S. This is defined and maximised in the following program:

$$\max_{\{t_{ij}\},\{s_{ij}\}} S = -\sum_{ij} t_{ij}[\log \frac{t_{ij}}{W_j^t} - 1] - \sum_{ij} s_{ij}[\log \frac{s_{ij}}{W_j^s} - 1] \quad , \qquad (2.55)$$

where $W_j^t$ and $W_j^s$ are the locational attractions of population zones j and service centres j respectively. Equation (2.55) is subject to the usual constraints on travel cost

$$\sum_{ij} t_{ij}c_{ij} = C^t \quad \text{and} \quad \sum_{ij} s_{ij}c_{ij}^s = C^s \quad , \qquad (2.56)$$

where $c_{ij}^t$, $c_{ij}^s$ are the costs of travel between i and j for workers and service users respectively and $C^t$ and $C^s$ are the associated mean travel costs. The economic base and origin constraints in equations (2.51) to (2.54) can be combined as

$$\sum_j t_{ij} - \beta \sum_k s_{ki} = (1-\beta)b_i \ , \quad \text{and} \qquad (2.57)$$

$$\sum_j s_{ij} - \sum_k t_{ki} = 0 \qquad (2.58)$$

where Coelho and Williams refer to equation (2.57) as the economic base constraint and (2.58) as a consistency condition although as will be clear below (2.58) is really another economic base constraint.

In maximising equation (2.55) the constraints in equations (2.56) to

(2.58) involve setting up the following Lagrangian multipliers: $\theta^t$ and $\theta^s$ for equations (2.56), $\xi$ and $\rho$ for equations (2.57) and (2.58) where these will be subscripted according to whether the variable enters the constraint in origin or destination form. The two models are derived as follows:

$$t_{ij} = W_j^t \exp\{-\xi_i + \rho_j - \theta^t c_{ij}^t\}, \quad \text{and} \tag{2.59}$$

$$s_{ij} = W_j^s \exp\{\rho_i - \beta\xi_j - \theta^s c_{ij}^s\} \quad . \tag{2.60}$$

The interlocking nature of the models in equations (2.59) and (2.60) is clearly displayed through the multipliers. The model can be solved and estimated simultaneously through direct optimisation of the objective function in equation (2.55) subject to its constraints in (2.56) to (2.58) However, the dual problem is an unconstrained optimisation problem of much reduced dimensionality and a more efficient procedure would involve minimising this dual.

A slightly more general case of this model results if exogenous population is included as in the model given in equations (2.22) to (2.25). Equation (2.53) now becomes

$$p_j = \gamma \sum_i t_{ij} + (1-\gamma)h_j \quad , \tag{2.61}$$

and this can be combined with equation (2.52) to give an alternative constraint equation to (2.58). Then

$$\sum_j s_{ij} - \gamma \sum_k t_{ki} = (1-\gamma)h_i \quad , \tag{2.62}$$

and the model which results from maximising S subject to equations (2.56), (2.57) and (2.62) is identical to that in equations (2.59) and (2.60) except that $\rho_j$ in equation (2.59) is replaced by $\gamma\rho_j$. The dual objective

function would show a greater difference in that $(1-\gamma)h_i$ would appear explicitly as well as $(1-\beta)b_i$.

Finally in this section, it only remains to indicate how this model can be set back into the linear framework. The model is not equivalent to the Lowry model because of the joint estimation of the spatial interaction models but it is structurally similar in that once these interaction model forms are known, the model is subject to the same constraints as the Lowry model. This is easily seen in linear terms. Noting then that

$$t_{ij} = e_i \frac{t_{ij}}{\sum_j t_{ij}} \quad \text{and} \quad s_{ij} = p_i \frac{s_{ij}}{\sum_j s_{ij}} ,$$

we can define

$$C_{ij} = \frac{t_{ij}}{\sum_j t_{ij}} \quad \text{and} \quad B_{ij} = \frac{s_{ij}}{\sum_j s_{ij}} . \qquad (2.63)$$

Using these definitions from equations (2.63) in the economic base relations, equations (2.53) and (2.54), we get the classic form

$$p_j = \sum_i e_i C_{ij} , \quad \text{and} \qquad (2.64)$$

$$e_j = \beta \sum_i p_i B_{ij} + (1-\beta)b_j \qquad (2.65)$$

which is equivalent to matrix equation (2.17). The same is true if the model with two inputs is considered, and this shows that the techniques of linear analysis used to enable empirical evaluation of the relevance of these relations can be used on such nonlinear models.

GENERALISED NONLINEAR LOWRY-LIKE MODELS.

We will now examine the complete two activity case where there is feedback

within the population and employment sectors. Define the probability of the population in i interacting with the population in j as $p_{ij}$ and the probability of employees in i interacting with the same in j as $e_{ij}$. These distributions are subject to origin constraints of the form

$$\sum_j p_{ij} = p_i \quad , \qquad \text{and} \tag{2.66}$$

$$\sum_j e_{ij} = e_i \quad . \tag{2.67}$$

The normalisation on $\{e_i\}$ and $\{p_i\}$ is as previously. Now from equation (2.26), the economic base relationships can be written out in elementwise form as

$$p_j = \alpha\sum_i p_{ij} + \gamma\sum_i t_{ij} + (1-\alpha-\gamma)h_j \quad , \qquad \text{and} \tag{2.68}$$

$$e_j = \beta\sum_i s_{ij} + \sigma\sum_i e_{ij} + (1-\beta-\sigma)b_j \quad . \tag{2.69}$$

Now noting that we have two additional origin constraints in equations (2.51) and (2.52), we have six constraints in all which can be reduced to four. There are various ways to effect this reduction and all are equivalent. Here we choose to show the self-dependence explicitly in each constraint by substituting the origin constraint directly into equations (2.68) and (2.69). Then the four constraints can be written as

$$\sum_j p_{ij} - \alpha\sum_k p_{ki} - \gamma\sum_k t_{ki} = (1-\alpha-\gamma)h_i \quad , \tag{2.70}$$

$$\sum_j s_{ij} - \alpha\sum_k p_{ki} - \gamma\sum_k t_{ki} = (1-\alpha-\gamma)h_i \quad , \tag{2.71}$$

$$\sum_j t_{ij} - \beta\sum_k s_{ki} - \sigma\sum_k e_{ki} = (1-\beta-\sigma)b_i \quad , \qquad \text{and} \tag{2.72}$$

$$\sum_j e_{ij} - \beta\sum_k s_{ki} - \sigma\sum_k e_{ki} = (1-\beta-\sigma)b_i \quad . \tag{2.73}$$

The other ways of representing these constraints are through removing equation (2.70) or (2.71), or (2.72) or (2.73) by substituting (2.70) into (2.71), or (2.72) into (2.73), or vice versa.

The group entropy function S to be maximised in this problem can be defined as

$$\max_{\substack{\{p_{ij}\},\{s_{ij}\} \\ \{t_{ij}\}\ \{e_{ij}\}}} \quad S = -\sum_{ij} p_{ij} \left[\log \frac{p_{ij}}{W_j^{\,p}} - 1\right] - \sum_{ij} s_{ij} \left[\log \frac{s_{ij}}{W_j^{\,s}} - 1\right]$$

$$-\sum_{ij} t_{ij} \left[\log \frac{t_{ij}}{W_j^{\,t}} - 1\right] - \sum_{ij} e_{ij} \left[\log \frac{e_{ij}}{W_j^{\,e}} - 1\right] , \qquad (2.74)$$

where $W_j^{\,p}$ and $W_j^{\,e}$ are population and employment attractors respectively. Two additional cost constraints reflecting the self-dependent interactions are required

$$\sum_{ij} p_{ij} c_{ij}^{p} = C^{p} \qquad \text{and} \qquad \sum_{ij} e_{ij}\, c_{ij}^{e} = C^{e} , \qquad (2.75)$$

where $c_{ij}^{p}$ and $c_{ij}^{e}$ are travel costs on route ij associated with each sector, and $C^{p}$ and $C^{e}$ are the respective mean travel costs. The problem then is to maximise S in equation (2.74) subject to travel cost constraints in equations (2.56) and (2.75), and the economic base constraints in equations (2.70) to (2.73). The Lagrangian multipliers $\theta^{t}$, $\theta^{s}$, $\theta^{p}$ and $\theta^{e}$ are associated with (2.56) and (2.75); multipliers $\phi$, $\rho$, $\xi$ and $\mu$ are associated with (2.70) to (2.73) where an appropriate origin or destination index is attached when optimisation occurs.

The four interaction models, which are the optimality conditions of the program just outlined, are

$$p_{ij} = W_j^p \exp\{-(1-\alpha)\phi_i + \alpha\rho_j - \theta^p c_{ij}^p\} \quad , \tag{2.76}$$

$$s_{ij} = W_j^s \exp\{-\rho_i + \beta(\xi_j + \mu_j) - \theta^s c_{ij}^s\} \quad , \tag{2.77}$$

$$t_{ij} = W_j^t \exp\{-\xi_i + \gamma(\phi_j + \rho_j) - \theta^t c_{ij}^t\} \quad , \qquad \text{and} \tag{2.78}$$

$$e_{ij} = W_j^e \exp\{-(1-\sigma)\mu_i + \sigma\xi_j - \theta^e c_{ij}^e\} \quad . \tag{2.79}$$

Estimation and solution can be achieved efficiently by minimising the unconstrained dual objective function, and its structure is of interest for the input variables enter the dual twice as two separate sets of cost terms. It is possible however that there are more parsimonious representations of the constraint set than that adopted here. Finally the model can be cast into its linear mould by defining

$$A_{ij} = \frac{p_{ij}}{\sum_j p_{ij}} \qquad \text{and} \qquad D_{ij} = \frac{e_{ij}}{\sum_j e_{ij}} \quad , \tag{2.80}$$

and using equations (2.63) and (2.80) in the model given earlier in equations (2.26) to (2.30).

Extensions to the multi-activity model are quite straightforward. As it is now regarded that this optimisation model framework is generally superior to the more *ad hoc* approach in which submodels are estimated separately, the linear framework is mainly of use in thinking about extensions to the model's causal structure and in empirical causal analysis, although is still of great use in developing insights into model solution and calibration. There are many extensions however to the use of optimisation theory and linear analysis in both theoretical and empirical contexts. The addition of other types of constraint, planning constraints for example, has been examined by Coelho and Williams (1978) as well as by Sharpe and Karlquist (1980). The optimisation of the model with respect to variables

other than interaction, particularly locational variables on the supply side, has been explored (Wilson, Coelho, Macgill and Williams,1981; Beaumont and Clarke, 1980), and Bertuglia and Leonardi (1980a, 1980b) have developed a variety of multi-activity versions using accessibility rather than entropy-maximising. Recently, the distinction between planning costs and benefits, and consumer surplus in terms of the objective functions used in deriving such models has been used to set the model in a game-theoretic context (Sharpe, Roy and Taylor, 1982). This is all in the spirit of unification of the field referred to at the outset of this review, and it is clear that a momentum has been established which has not yet worked itself out in any sense (Batty, 1981b; Brotchie, Dickey and Sharpe, 1980).

CONCLUSIONS.

A broad framework in which all the subsequent analysis in this thesis which concerns model solution dynamics and spatial variation due to model structure has been established, and this will guide the various themes which will be developed from now on. As these themes are intricately related, the framework of this chapter is of use in emphasising relationships between later chapters. In the sequel, calibration methods and simulation techniques will be dealt with extensively but it is perhaps appropriate to end this chapter with some speculation on the field in general, rather than this thesis in particular. This is particularly apt as this review chapter was written last in the thesis and thus has been written in the light of the research in all subsequent chapters.

Although considerable progress has been made in urban modelling in the last twenty years, this has been mainly in methodological areas such as

those developed in this thesis. What are now required are applications building and refining ideas such as those presented here. In the next decade it is likely that attention will veer towards more substantive questions, and issues concerning what is to be modelled, rather than modelling methods. In a sense, one of the great disappointments of modelling practice has been the inability of theorists to suggest model structures which capture the qualitative flavour of urban systems and problems. The systems being modelled are fairly similar to those of significance twenty years ago, despite major changes by planners and policy-makers over what they consider to be of current importance.

Perhaps the major problem facing the field now, however, concerns the dearth of empirical applications. In the social sciences, theory always proceeds ahead and somewhat independently of practice although for real progress to be made, practice and theory must frequently meet and gell. Here for example, the importance of the linear extensions to the Lowry model and the emphasis on the choice of appropriate structures can only be complete when these ideas are used empirically. This has rarely happened, and the synthesis involving optimisation which would have been useful a decade or more ago when such models were being developed practically, is now of mainly theoretical interest. Yet the advances in the field have been so great that there is now the real prospect that these techniques could be used to design and apply urban models which perform more sensitively and appropriately than similar models did a decade ago. If so, such models will have much greater predictive power with all the consequences for urban planning. Only by extensive but careful, considered and technically sophisticated applications will the promise of these advances be borne out and the next decade should be focussed on developing progressive practice.

# CHAPTER 3.

## A THEORETICAL FRAMEWORK FOR PSEUDO-DYNAMIC URBAN MODELS.

In reviewing contemporary developments in urban modelling in the last chapter, it was argued that greatest progress has been made in the development of cross-sectional static models, rather than dynamic. Principles for handling the dimension of real time, and the establishment of relevant mechanisms for representing processes of urban change have been difficult to research, and existing dynamic models exhibit a degree of arbitrariness which is disturbing. The widely known *Urban Dynamics* model (Forrester, 1969; Alfeld and Graham, 1976), for example, is really no more than a demonstration that a dynamic treatment of urban phenomena is required, for the urban system and its behaviour through time which is the subject of the model, is hypothetical.

In contrast, the dynamic version of the Access and Land Development model (Schneider, 1976) although based on a theory of the urban system which is intuitively acceptable and in part, empirically known, uses a mathematical framework based on the Lotka-Volterra equations, which is specified arbitrarily. The more recent development of urban models based on embedding spatial interaction models into a similar frameworks which emphasise catastrophe, bifurcation and fluctuation (Wilson, 1981) are also problematic

in the same regard although the work of Varaprasad and Cordey-Hayes (1982) shows considerable potential in such developments. Furthermore, it has also been difficult to capture the kinds of relationships which characterise cross-sectional static models in truly dynamic form, and this has led to dynamic models which lack the richness of the static models which they seek to improve, and perhaps replace.

To avoid these problems even in a partial way, it appears that a framework is required in which static and dynamic models exist at opposite ends of a continuum involving the treatment of time. In this sense, a static model would have a dynamic equivalent and vice versa, and one could be derived from the other by aggregation or disaggregation of the appropriate time dimension. In fact, it is quite easy to suggest frameworks which reflect this idea; the simplest would be one in which static models could be made dynamic by simply indexing the variables according to time. Therefore, for such a framework to be other than trivial, an additional organising principle is required. Usually only static and dynamic forms of model identify the ends of the time continuum, and any intermediate form of model can be identified with one end or the other: comparative static with static, quasi-dynamic with dynamic and so on. But if a third model form is identified which contains both static and dynamic elements, then the framework must be specified at a higher level of complexity to embrace such a form. This third form will be referred to as a *pseudo-dynamic* model, thus representing an intermediate position between fully static and fully dynamic models. Because such a model contains both static and dynamic elements, this implies that two or more time streams characterise the framework and by aggregation or disaggregation of the appropriate stream, fully static and dynamic models can be derived. In such a framework, static models will contain implicit time dimensions

whereas dynamic models will contain explicit ones, while pseudo-dynamic models will contain both implicit and explicit time streams.

IDEAS CONCERNING PSEUDO-DYNAMICS.

Chapters 3 to 9 of this thesis will be broadly concerned with postulating, elaborating and applying the idea of a pseudo-dynamic model, and this chapter will be specifically orientated towards the dynamic framework through which such models can be derived. In Chapters 4 and 5, more detailed forms of pseudo-dynamic model will be explored and an attempt will be made to calibrate such a model to a real situation. Chapters 6 to 9 will be concerned with more fundamental issues of calibration involving analogies with the optimal control of a dynamic system and with matrix iterative analysis but the emphasis in all these chapters will be upon generating new insights concerning the operational development of both static and dynamic urban models. As this is to be accomplished through the device of the pseudo-dynamic model, it is worthwhile discussing the meaning attached to such a model before it is formally introduced.

Consider the class of dynamic processes which operate through time other than historical time: for example, models in which solutions are reached iteratively through trial and error elimination reflect simultaneous relationships which have to be solved sequentially. These models might be regarded as pseudo-dynamic if the sequential solution procedure implies a kind of historical time through which the system is changing. On the other hand, there are models which are characterised by different types or 'streams' of historical time; for example, the actual system time might be distinguished from the time when activity was first generated, and if one of these time streams were to be collapsed, the model would be

pseudo-dynamic. There are many other examples to be found. Multiplier models in which the multiplier effects pertain to historical time, but are worked out in terms of model solution time, models which are static in nature but are 'artificially' grown to a cross-section in time using some dynamic process, these are the types of models which are prime candidates for treatment in pseudo-dynamic terms. Indeed, it might be argued that these models involve approximations to historical time through a concept of pseudo-time, and an essential first step is to describe the framework in which this historical time is explicit.

The dynamic framework to be outlined here, meets the requirements posed for a fully dynamic process by several researchers in the field of urban modelling (Sayer, 1975; Williams and Wilson, 1977) and time will be clearly involved 'in an explicit and essential way' (Curry and MacKinnon, 1975; Samuelson, 1948). The dynamic framework is fashioned in fairly well-defined terms, and as its equilibrium properties are well-known, no emphasis will be laid on proving the existence and uniqueness of equilibrium. Rather the emphasis will be on exploring the various types of process which characterise the system, and the ways in which these processes unfold through time. Furthermore, the focus will be upon presenting operational models, and thus the equation systems given below will have a numerical flavour. One essential argument relates to the idea that new insights into existing models can be gained by setting up more general frameworks within which existing models can be cast. Indeed, in later chapters some substance will be given to this notion when existing models are reinterpreted, and new methods of calibrating these models

are derived from pseudo-dynamic considerations. Finally, it is possible to derive new model forms from this type of analysis, and although new forms will not be extensively discussed, the next chapter will demonstrate the potential of the framework in this regard.

In the development of the dynamic framework, the general dynamic model will be first stated in highly aggregate terms. Two types of process characterise the model, the first based on the generation of activity endogenous to the model through time, and the second based on the location of that same activity to zones of a bounded urban region. The generation of activity is treated first and the various processes involving new change and changes in existing activity are described. The location process also has a dynamic quality in that locational attractions are lagged through time. Appropriate location models are derived by an information- theoretic method used previously by the author (Batty and March 1978), and then the complete model is assembled.

The derivation of a pseudo-dynamic form is accomplished in two stages. First, a closed form for the dynamic model is derived, and second, this is aggregated in various ways to derive pseudo-dynamic models. This chapter ends with a statement of one such model which forms the starting point of the next two chapters where it is explored and applied. Although somewhat removed from the immediate concerns of the present chapter, Chapters 6 to 9 involve an explicit use of the pseudo-dynamic process in deriving estimation procedures appropriate for certain existing operational models. But first the

dynamic model from which all else is derived, will be presented.

## GENERAL STRUCTURE OF THE DYNAMIC URBAN MODEL.

In its simplest form, the dynamic model predicts the amount and location of two related activities from information concerning some exogenous input activity. For example, population and service employment can be predicted from basic employment and the way in which this is modelled, is based upon the sequential generation of each endogenous activity from the input: that is, population is generated from basic employment, services are generated from population, more population is generated from services and so on. In principle, this kind of process can be extended to any number of endogenous and exogenous sectors, and the model need not be restricted to population and employment activities. The structure is completely general and as long as there is meaning to the process of sequential generation, there is no reason why the model should not be applicable to any socio-economic, perhaps even physical system.

However, specific applications will depend on whether or not the structure can be meaning-fully applied to the system of interest, and here the model will be based on the simple distinction into the population and employment sectors of an urban system. Disaggregation into different subsectors is easy to accomplish but it would add nothing to the logic, and as it would give rise to a superficial complexity which may divert the reader from the essential argument, it has been avoided. The equations which follow are already complicated and somewhat cumbersome without the addition of further

detail. Note that in the following presentation, as one endogenous activity is directly and simply related to the other, the forms of the state equations governing the configuration of the system at any point in time are similar for each activity. Thus usually only one set of equations need be presented, for the other set follows immediately. Whenever this occurs, the reader will be forewarned.

Another essential characteristic of the dynamic framework is the treatment of two types of change: change due initially to changes in the exogenous activity, and change due to changes in existing activity. The first type of change is the easiest to handle and is referred to as new change; it is the direct result of changes in the input which are 'new' in each period. Note that such change may be positive or negative, implying growth or decline. The second type of change relates to changes in the existing stock of activity, and whereas new change leads to different levels of total activity in the system, changes in the existing stock only redistribute what is there already.

At each point in time, the existing stock can be divided into 'movers' and 'stayers' which denote that activity which has redistributed itself in a previous time period and that which remains stable, that is, identical to its previous distribution. The movers are notated using superscript m and the stayers superscript s. So at each point in time, activity is composed of movers, stayers, and new change. In the following presentation, two sets of state equations are defined for employment and population. Net change in activity x in any time period [t:t-1] is notated by $\Delta x(t)$, and gross change, that is changes

in various components making up total change by $\Delta^*x(t)$ or $\Delta'x(t)$. The lower case bold[†] letters all refer to 1 x N row vectors and the bold capitals to N x N matrices, where N is the total number of zones in the urban system. Note that it is assumed without loss of generality, that there is population and employment in each zone.

Employment $\underline{e}$ (t) at any time t is calculated from

$$\underline{e}(t) = \underline{e}^s(t) + \underline{e}^m(t) + \Delta^*\underline{e}(t), \qquad (3.1)$$

where $\underline{e}^s(t)$, $\underline{e}^m(t)$ and $\Delta^*\underline{e}(t)$ are vectors of the stayers, the movers and the new change in employment occurring in the time period [t:t-1]. The new change is made up of changes in basic employment (the input) $\Delta\underline{b}(t)$ and changes in service employment $\Delta^*s(t)$ which arise directly from changes in the input. Then

$$\Delta^*\underline{e}(t) = \Delta^*\underline{s}(t) + \Delta\underline{b}(t). \qquad (3.2)$$

As it is assumed that basic employment is entirely exogenous to the system in that it provides the driving force for new change, if it is to be redistributed, this must be achieved exogenously to maintain consistency. Therefore only service activity can be redistributed in any time period, and this implies that the mover-stayer components refer exclusively to services. Thus

$$\underline{e}^s(t) = \underline{s}^s(t) + \underline{b}(t-1), \quad \text{and} \qquad (3.3)$$

$$\underline{e}^m(t) = \underline{s}^m(t), \qquad (3.4)$$

where $\underline{s}^s(t)$ and $\underline{s}^m(t)$ are the service stayers and movers in the period [t:t-1] and $\underline{b}(t-1)$ is the total basic employment existing at t-1. Substituting for $\Delta^*\underline{e}(t)$, $\underline{e}^s(t)$ and $\underline{e}^m(t)$ in equation (3.1) from

---

[†]Bold symbols are indicated by underlining throughout the text.

equations (3.2), (3.3) and (3.4) respectively gives

$$\underline{e}(t) = \underline{b}(t-1) + \underline{s}^{s}(t) + \underline{s}^{m}(t) + \Delta^{*}\underline{s}(t) + \Delta\underline{b}(t). \qquad (3.5)$$

It is assumed that population is derived from employment, and thus an analogous equation exists for population but without any explicit exogenous influences:

$$\underline{p}(t) = \underline{p}^{s}(t) + \underline{p}^{m}(t) + \Delta^{*}\underline{p}(t), \qquad (3.6)$$

where $\underline{p}(t)$, $\underline{p}^{s}(t)$, $\underline{p}^{m}(t)$ and $\Delta^{*}\underline{p}(t)$ are vectors of total population at time t, stayers, movers and new population change in the period [t:t-1] respectively. The way in which equations (3.5) and (3.6) relate to one another is quite complicated and will be discussed as a separate process in a later section.

PROCESSES OF URBAN CHANGE.

The endogenous variables, population and service activity, are the result of complex patterns of generation and in the case of both new change and change in the existing stock, changes occurring in any time period [t:t-1] can be the result of original changes in some earlier time period. In other words, change in the period [t:t-1] can be made up of a series of components each originating at previous time periods, and to capture this spectrum of change, it is necessary to develop a more elaborate notation. The simplest type of change is new change: for example, the change in activity x at time $\tau$ can be traced back to time z which is its initial point of origin, that is, changes in activity at time z are still working themselves out

at time $\tau$. This type of change can be notated as $x(\tau,z)$ where $\tau \geqslant z$. These two time streams $\tau$ and z relate, in this model, to changes in the service activity and population directly associated with changes in the input, basic employment.

However, it is necessary to say something about the form of this generation process for it is unlikely to continue indefinitely, A reasonable assumption is that the amount of change generated at time $\tau$ originally associated with z, will get less as $\tau$ gets greater. In other words, the effect of the original change at z will die away as more time elapses from z. For the case of service activity $s(\tau,z)$, (population follows an identical process), it is postulated that

$$\underline{s}(z,z) > \underline{s}(z+1,z) > \underline{s}(z+2,z) > \ldots > \underline{s}(\tau,z),$$

and it is assumed that the last significant change $s(\tau,z)$ occurs T+1 time periods after the original change in time period [z:z-1], that is, in the period [z+T:Z+T-1]. Thus the change $\underline{s}(z+T+1,z)$ is assumed to be zero, and the process of generation has a life of T time units.† Noting that the total length of significant lag in origin time is T units, then in any time period [t:t-1], the spectrum of new change is given as

† The index T is also used more generally in chapters 3 to 9 to indicate some future point in time. Wherever this so, the difference between its use in identifying the life of the generation process, and its more general usage, will be made obvious. Note also that T is occasionally used to define trips and as the matrix transpose operator.

$$\left.\begin{aligned} \Delta^*\underline{s}(t) &= \sum_{z=t-T}^{t} \underline{s}(t,z), \qquad \text{and} \\ \Delta^*\underline{p}(t) &= \sum_{z=t-T}^{t} \underline{p}(t,z) \quad , \end{aligned}\right\} \quad (3.7)$$

where $\underline{s}(t,z)$ and $\underline{p}(t,z)$ are the services and population activities generated at t, originating at z. Note that it is not necessary to use the difference operator $\Delta$ for the time notation is sufficiently detailed a means of indexing this change.

The situation is more complicated for changes in the existing stock for it is necessary to distinguish between the time when the activity is moved $\tau$, the time when the activity was first generated w, and the time of origin when the generation of this activity was first initiated z. At present, it is only necessary to consider movers because stayers can always be defined by comparing movers with the previous configuration of the system and thus the variable $x(\tau,w,z)$ refers to movers in time $\tau$, who originally entered the system in time w, based on a process of generation originating at time z. To give substance to these notions, it is necessary to explore the change in existing activity in more detail. Because activity has to exist before it can move, then this implies that the minimum length of time from when activity is first generated to when it is able to move is one time period. In this model, it is thus assumed that activity is first able to move one time period after it has been generated. Therefore, $\tau > w \geqslant z$. Just as new change is a spectrum of change originating at previous time periods, movers in any time period [t:t-1] relate to those generated at previous times w originating at previous times z, noting that $t > z$.

As in the case of new change, it is assumed that movers redistribute themselves according to a similar process of generation (regeneration). For the case of service movers $\underline{s}^m(\tau,w,z)$ (population movers follow an identical process), this assumption means that

$$\underline{s}^m(t,z,z) > \underline{s}^m(t+1,z+1,z) > \underline{s}^m(t+2,z+2,z) > \ldots > \underline{s}^m(\tau,w,z)$$

where the origin lag between z and w is no greater than T. Then the process in which activity originally moves at time t dies away until at time t+T, the last significant change in the movers from time t is recorded. However, activity which exists, generates a potential move at every subsequent time period and thus the spectrum of change which characterises the movers in any time period [t:t-1] is made up of movers who originate at *all* previous times z and are generated at *all* previous times w.

In summing the variable $x(\tau,w,z)$ over z, the range of summation is from the beginning of time z=0 which is purely notional as will become clear later, to z=t-1, t>z (the one period mover lag). The range of w is from z to z+T where T is the total life of the generation process. One final complication exists: because new activity originating in the period [t-1:t-T-1] has not yet fully worked itself out, then it is necessary to separate out this activity from activity originating prior to time t-T. Then the services and population movers in the period [t:t-1] are calculated from

$$\left.\begin{aligned} \underline{s}^m(t) &= \sum_{z=0}^{t-T-1} \sum_{w=z}^{z+T} \underline{s}^m(t,w,z) + \sum_{z=t-T}^{t-1} \sum_{w=z}^{t-1} \underline{s}^m(t,w,z), \text{ and} \\ \underline{p}^m(t) &= \sum_{z=0}^{t-T-1} \sum_{w=z}^{z+T} \underline{p}^m(t,w,z) + \sum_{z=t-T}^{t-1} \sum_{w=z}^{t-1} \underline{p}^m(t,w,z) , \end{aligned}\right\} \quad (3.8)$$

where $\underline{s}^m(t,w,z)$ and $\underline{p}^m(t,w,z)$ are the service and population movers originating at z, generated at w and moving at t.

A similar type of equation based on the equation for movers could be developed for stayers but as this is also a function of previous movers and stayers, its form can be postponed until the more detailed investigation of the model's generation processes is presented in a later section. It is possible, however, to define the general structure of the model in terms of the ideas introduced so far, and to state the overall system constraints which must be met. Then the change in employment which is made up of movers, new services and new basic employment, called $\Delta'\underline{e}(t)$ is given by

$$\Delta'\underline{e}(t) = \sum_{z=0}^{t-T-1} \sum_{w=z}^{z+T} \underline{s}^m(t,w,z) + \sum_{z=t-T}^{t-1} \sum_{w=z}^{t-1} \underline{s}^m(t,w,z) + \sum_{z=t-T}^{t} \underline{s}(t,z) + \Delta\underline{b}(t), \tag{3.9}$$

subject to the following constraint on the total level of service stayers

$$\sum_i s_i^s(t) = \sum_i e_i(t-1) - \sum_i s_i^m(t) - \sum_i b_i(t-1), \quad i=1,2,\ldots,N.$$

The same type of equation can be developed for population, and the gross change $\Delta'\underline{p}(t)$ made up of movers, and new population is given by

$$\Delta'\underline{p}(t) = \sum_{z=0}^{t-T-1} \sum_{w=z}^{z+T} \underline{p}^m(t,w,z) + \sum_{z=t-T}^{t-1} \sum_{w=z}^{t-1} \underline{p}^m(t,w,z) + \sum_{z=t-T}^{t} \underline{p}(t,z), \tag{3.10}$$

subject to the following constraint on the total level of population stayers

$$\sum_i p_i^s(t) = \sum_i p_i(t-1) - \sum_i p_i^m(t) , \qquad i=1,2,..,N.$$

Note that the constraints are purely accounting equations which will be met if the movers are modelled consistently and if the process of calculating the stayers is related to the known patterns of previous movers and stayers predicted by the model.

The model stated in equations (3.9) and (3.10) is at a highly aggregated level and the way in which the activities relate to one another through the processes of generating new change and regenerating change in the existing stock now need to be described. To do this, the simplest of the processes - that characterising new change will be described first, and this will help in demonstrating the process involved in redistributing movers. Lastly, the process of modelling the stayers will be treated, and as this depends upon the movers and new change, it is the most complicated.

MODELS FOR GENERATING NEW URBAN CHANGE.

It has already been implied that the causal structure of the model depends upon generating endogenous variables in sequence starting from the initial input, and it seems appropriate that the strength of these sequences should become weaker as time elapses from the initial input which starts the process. Consider the case of employment and population: basic employment generates basic population which requires services. Services generate their own dependent population which also require to be serviced and so on. The process is well-known: it exists as the economic-base method in regional economics, the input-output model of general equilibrium theory and the multiplier process in Keynesian economics. But it is of much wider significance in that it is a model

of any system in which cause and effect can be characterised in this uni-directional way.

In this context, the model is organised around the economic base mechanism where basic employment is taken as the input and service and population activity as the output. This gives rise to models which are well-known, for example those of the Lowry (1964) genus, and it also means that the models presented in this part of the thesis can be always made operational. The process of generation from an initial input vector of basic employment $\Delta\underline{b}(t)$ is as follows: $\Delta\underline{b}(t)$ generates $\underline{p}(t,t)$ which in turn generates $\underline{s}(t,t)$. In the next time period [t+1:t], s(t,t) generates $\underline{p}(t+1,t)$ and this in turn generates $\underline{s}(t+1,t)$. For the process to be meaningful, it must converge, that is, $\underline{s}(\tau+1,t) < \underline{s}(\tau,t)$ and $\underline{p}(\tau+1,t) < \underline{p}(\tau,t)$. As stated above, the process is assumed to have converged after T+1 time periods have elapsed from the initial change in basic employment in [t:t-1].

The length of the process T could of course be dependent upon t or it could follow some random pattern over time. The decision to phase the sequential generation of endogenous activity into fixed time intervals over the whole process, is only one of convenience, and this could also be varied if required as long as the process converged in some sense. Variation in time lags could be easily incorporated into the model, but the assumption of a fixed lag and length of process does not involve any loss of generality, and it considerably simplifies the ensuing algebra. Readers who are interested in exploring the effects of varying these fixed intervals are referred to Bartholomew

(1982) for a discussion of such effects in similar linear models of the Markov type.

The way in which activities are derived from one another is of considerable importance in this model because it is here that relationships over space can best be dealt with. There are two major relationships between activities such as employment and population, and vice versa, and these pertain to scale and distribution. Scale relationships such as activity rates can be incorporated but the way in which distribution is handled is more complex. The obvious way to derive one vector from another is through a matrix of relationships, and in the case of spatial activities, such matrices would summarise spatial interactions. In the above process, population $\underline{p}(\tau,t)$ can be derived from employment $\underline{s}(\tau-1,t)$ using an interaction matrix $\underline{A}(\tau,t)$ which records the relation or interaction between any employment location i and population location j. Typically, this might relate to the journey to work, but it also includes a scaling effect to convert employment into population.

In deriving $\underline{s}(\tau,t)$ from $\underline{p}(\tau,t)$, another matrix $\underline{B}(\tau,t)$ is required which scales population to services, and summarises the spatial demands by the population in j for services in k. Such demands might be measured by shopping trips, business transactions to the home and so on. The process of generating population and services can now be summarised: note that the matrices are square (N x N) but this too is only an assumption of convenience. Then the process is

$$\underline{p}(t,t) = \Delta\underline{b}(t)\underline{A}(t,t), \tag{3.11}$$

$$\underline{s}(t,t) = \underline{p}(t,t)\underline{B}(t,t), \tag{3.12}$$

$$\underline{p}(t+1,t) = \underline{s}(t,t)\underline{A}(t+1,t), \quad \text{and} \tag{3.13}$$

$$\underline{s}(t+1,t) = \underline{p}(t+1,t)\underline{B}(t+1,t). \tag{3.14}$$

Recurrence on equations (3.13) and (3.14) to time t+T leads first to the population relation

$$\underline{p}(t+T,t) = \Delta\underline{b}(t) \prod_{\tau=t}^{t+T-1} \underline{A}(\tau,t)\underline{B}(\tau,t)\underline{A}(t+T,t), \tag{3.15}$$

and then to the employment relation

$$\underline{s}(t+T,t) = \Delta\underline{b}(t) \prod_{\tau=t}^{t+T} \underline{A}(\tau,t)\underline{B}(\tau,t). \tag{3.16}$$

In the dynamic model, it is assumed that activity is being generated from new inputs which occur in every time period. Basic employment $\Delta\underline{b}(z)$ is input at every time z and thus the repercussions from these inputs in terms of population $\underline{p}(\tau,z)$ and service employment $\underline{s}(\tau,z)$ will occur for T+1 time periods from the initial input. Thus at any period in time, new activity change will have originated at the previous T+1 periods. This process of generation is graphed in Figure 3.1 where the horizontal axis of the chart refers to the time of generation and the vertical axis to the time of origin. Clearly at any period $[\tau:\tau-1]$, there are T+1 components of change which constitute the spectrum. The appropriate equations for total population and service employment change in $[\tau:\tau-1]$ can be derived by summing equations (3.15) and (3.16) over t. In this instance, the index z is used to indicate origin time and t generation time. Then

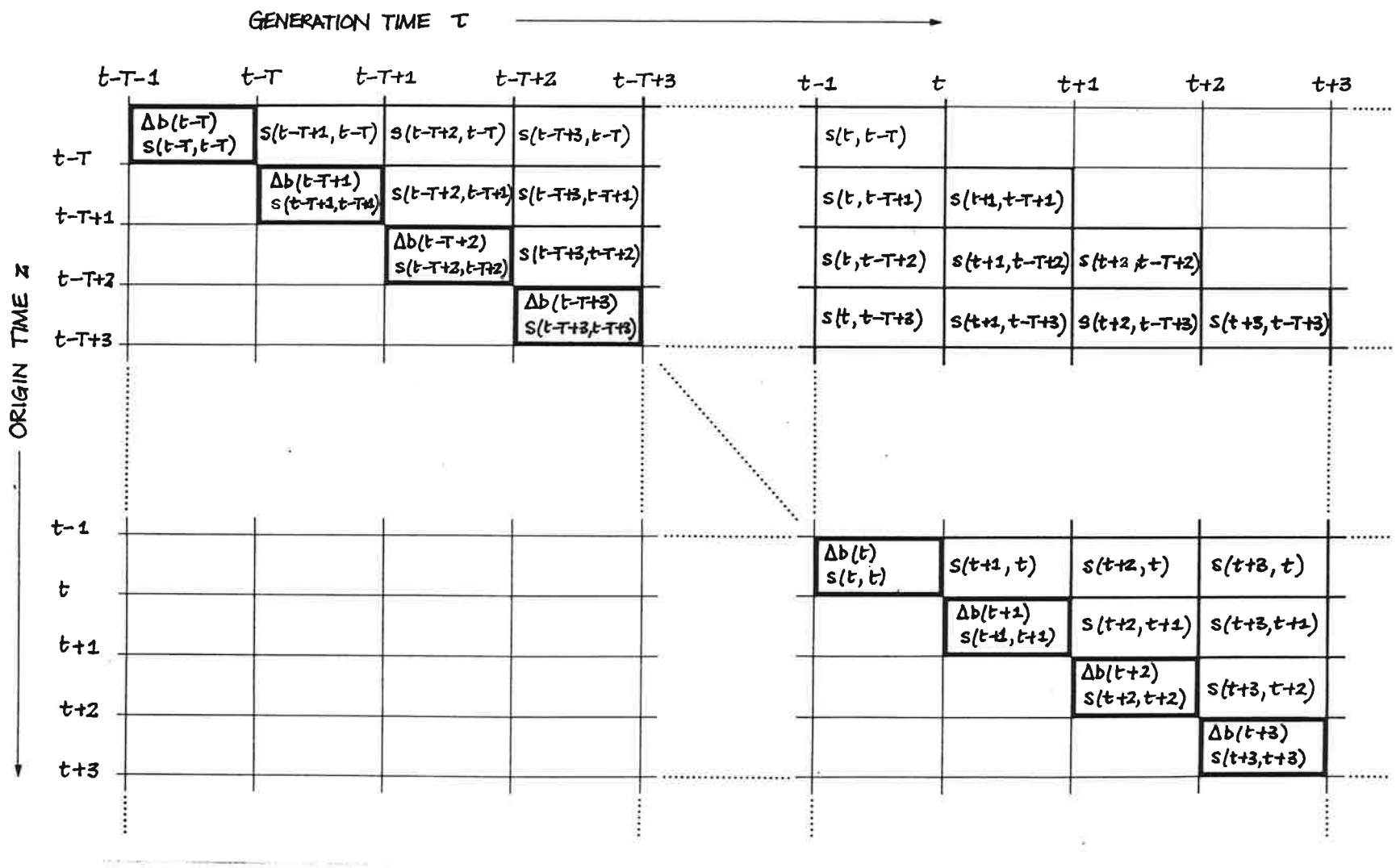



Figure 3.1: The Generation of New Activity Streams.

$$\sum_{z=t-T}^{t} \underline{p}(t,z) = \sum_{z=t-T}^{t} \Delta\underline{b}(z) \prod_{\tau=z}^{t-1} \underline{A}(\tau,z)\underline{B}(\tau,z)\underline{A}(t,z), \tag{3.17}$$

$$\sum_{z=t-T}^{t} \underline{s}(t,z) = \sum_{z=t-T}^{t} \Delta\underline{b}(z) \prod_{\tau=z}^{t} \underline{A}(\tau,z)\underline{B}(\tau,z). \tag{3.18}$$

For equations (3.17) and (3.18) to be meaningful, it is clear that the process time T must be so defined that the matrix product converges to the null matrix $\underline{0}$. Then as in any convergence, a limit matrix $\underline{\varepsilon}$ is defined such that

$$\prod_{\tau=t-T-1}^{t} \underline{A}(\tau,z)\underline{B}(\tau,z) < \underline{\varepsilon} \quad ,$$

and this defines T. There are several ways of achieving this and to conclude this section, it is worth mentioning the more obvious.

The matrices $\underline{A}$ $(\tau,z)$ and $\underline{B}(\tau,z)$ relate scale to distribution (interaction) effects, and it is likely that the scale effects will control the convergence of the process if appropriately sepcified. Assume that the scale effects can be represented by matrices $\underline{\Lambda}(z)$ and $\underline{\Gamma}(z)$ associated with $\underline{A}(\tau,z)$ and $\underline{B}(\tau,z)$ respectively, and it is clear that these are dependent only upon the origin time z which is fixed for each process. Therefore scale and distribution effects are separable and combine as

$$\underline{A}(\tau,z) = \underline{T}(\tau,z)\underline{\Lambda}(z), \quad \text{and} \tag{3.19}$$

$$\underline{B}(\tau,z) = \underline{S}(\tau,z)\underline{\Gamma}(z) \ . \tag{3.20}$$

The matrices $\underline{T}(\tau,z)$ and $\underline{S}(\tau,z)$ relate to the pattern of spatial interactions associated with service-population and population-service relationships respectively. If it is assumed that the scale effect relating

one activity to the other is constant over space as well, then $\underline{\Lambda}(z)$ and $\underline{\Gamma}(z)$ are diagonal scalar matrices with constants $\lambda(z)$ and $\gamma(z)$ on the respective main diagonals. Note that $\lambda(z)$ and $\gamma(z)$ can be regarded as the inverse activity-rate and population-serving ratio respectively, calculated for the whole system.

Forming the matrix product from $\tau=t$ to $t+T$ for the matrices $\underline{A}(\tau,z)$ and $\underline{B}(\tau,z)$ leads to the following simplification

$$\prod_{\tau=t}^{t+T} \underline{A}(\tau,z)\underline{B}(\tau,z) = [\underline{\Lambda}(z)\Gamma(z)]^{T+1} \prod_{\tau=t}^{t+T} \underline{I}(\tau,z)\underline{S}(\tau,z). \qquad (3.21)$$

A sufficient condition for convergence would require that the matrix $[\underline{\Lambda}(z)\underline{\Gamma}(z)]^T$ converge to the null matrix at $T\to\infty$. This in fact is the basis of convergence of the economic base process, and it means that $\lambda(z)\gamma(z)< 1$. For a nontrivial economic base process, this must be true in order that the process generate finite values of population and service employment from basic employment. It is also possible that this condition be satisfied if $\lambda(z)$ and $\gamma(z)$ depend upon $\tau$ as well; in fact, Goldner's (1974) PLUM (Projective Land Use Model) uses this idea, although there are problems in generating the correct total level of activity which have to be resolved by iteration. Finally, it is worth noting that this dynamic process has already been used as the basis for a simple dynamic model developed by the author for the Reading subregion (Batty, 1976).

THE GENERATION OF MOVERS.

The mover variables $\underline{s}^m(t,w,z)$ and $\underline{p}^m(t,w,z)$ were defined above and the

purpose of this section is to postulate models which simulate their process of redistribution through time. In fact, the model is analogous to that used in generating and allocating new change, which in this case, is a dynamic economic base model. It is necessary to use such a model to maintain consistency between the way in which change is first allocated, and then reallocated, but more important is the fact that the process seems reasonably realistic. If it is assumed that service activity and population is built up from the pattern of basic employment, then it seems likely that movers originating from basic population set off a sequence of moves through dependent population and services. Furthermore if the initial pattern is derived using an economic base type process, any change within this pattern must be related to the original pattern to enable consistent accounting. To explore this process, it is only necessary to examine service movers because population is directly related to services in the way already described. Here, as in the previous section, a model of one set of moves from activity originating at time t, will first be presented, and then this will be extended to a series of migration streams which include activity which has originated at all significant previous periods of time.

Consider the services originating from the basic employment input at time t. These services are generated in future time periods up to t+T, and exist as $\underline{s}(t,t)$, $\underline{s}(t+1,t)\ldots\underline{s}(t+T,t)$. Now in time period t+1, the previously generated services $\underline{s}(t,t)$ form part of the mover pool, the pool of activity which is potentially able to relocate. It is assumed that a fixed proportion $\underline{0} \leq \underline{\alpha}(t+1,t+1) \leq \underline{I}$ of these services relocate and these new service movers $\underline{s}^m(t+1,t,t)$ are found by applying

this mover ratio matrix to the initial input $\Delta\underline{b}(t)$ and reallocating this proportion using new scale-distribution matrices $\underline{\tilde{A}}(t+1,t)$ and $\underline{\tilde{B}}(t+1,t)$. These movers are given by

$$\underline{s}^m(t+1,t,t) = \Delta\underline{b}(t)\underline{\alpha}(t+1,t+1)\underline{\tilde{A}}(t+1,t)\underline{\tilde{B}}(t+1,t). \qquad (3.22)$$

Note that $\underline{\alpha}(t+1,t+1)$ is an N x N diagonal matrix of mover ratios which may vary zonally; here the later analysis is helped if it is assumed that $\underline{\alpha}(t+1,t+1)$ is scalar diagonal (constant $\alpha(t+1,t+1)$ for each zone i). The process initiated in equation (3.22) sets off a series of repercusssions through time in which the first service movers at t+1 generate more service moves at t+2 and so on. Then

$$\underline{s}^m(t+2,t+1,t) = \underline{s}^m(t+1,t,t)\underline{\tilde{A}}(t+2,t)\underline{\tilde{B}}(t+2,t), \qquad (3.23)$$

and by recursion on equations (3.22) and (3.23) to t+T, the general mover relation is derived as

$$\underline{s}^m(t+T,t+T-1,t) = \Delta\underline{b}(t)\underline{\alpha}(t+T,t+1) \prod_{\tau=t+1}^{t+T} \underline{\tilde{A}}(\tau,t)\underline{\tilde{B}}(\tau,t). \qquad (3.24)$$

The characteristics of this mover process are assumed to be similar to the process of generating new change. It is assumed that the life of the process is T units, that is, the last movers associated with the initial moves at t+1 are generated in t+T+1. Furthermore, the process is assumed to have converged at this point, and it is a requirement of the model that the existing level of activity is preserved through the matrices $\underline{\tilde{A}}(\tau,t)$ and $\underline{\tilde{B}}(\tau,t)$. This will be discussed in more detail later.

This process can be generalised in two main ways: first it is clear that if the original activity generated at time t gives rise to a

process of movement starting at time t+1, then the same activity has the potential for starting a sequence of moves at any time $\tau > t$. In other words, once activity exists within the system, a mover ratio is applicable in every time period subsequent to the time when it was first generated. The second generalisation relates to the fact that activity which has been originally generated at *any* period prior to time t is able to set off a sequence of moves which have the form of equation (3.24).

This kind of complexity in time streams is illustrated by the three-dimensional Lexis-like diagram in Figure 3.2 (Rees and Wilson, 1977). The three dimensions relate to the three critical time streams characterising movers: the time z when the activity originated in the system, the time w when the activity was first generated, and the time v when the activity set off the process of moves. As each mover process has a life of T units, there are at any one point in time T+1 mover streams associated with any activity originating at z. This is clear by examining the v-w face of Figure 3.2 for any given time t. It also means that for any time period [t:t-1], only mover ratios from t to t-T need be considered in modelling movers, because ratios prior to t-T are no longer generating significant movers. The equation describing a move at time t is thus

$$\underline{s}^{m}(t,w,z) = \Delta\underline{b}(z)\underline{\alpha}(t,v) \prod_{\tau=v}^{t} \underline{\tilde{A}}(\tau,z)\underline{\tilde{B}}(\tau,z), \quad t-T \leq v \leq t, \quad 0 \leq z \leq t-1. \tag{3.25}$$

The index v is related to the three time dimensions by $v=t-w+z$, and it is immediately clear that its range is t-T to t.

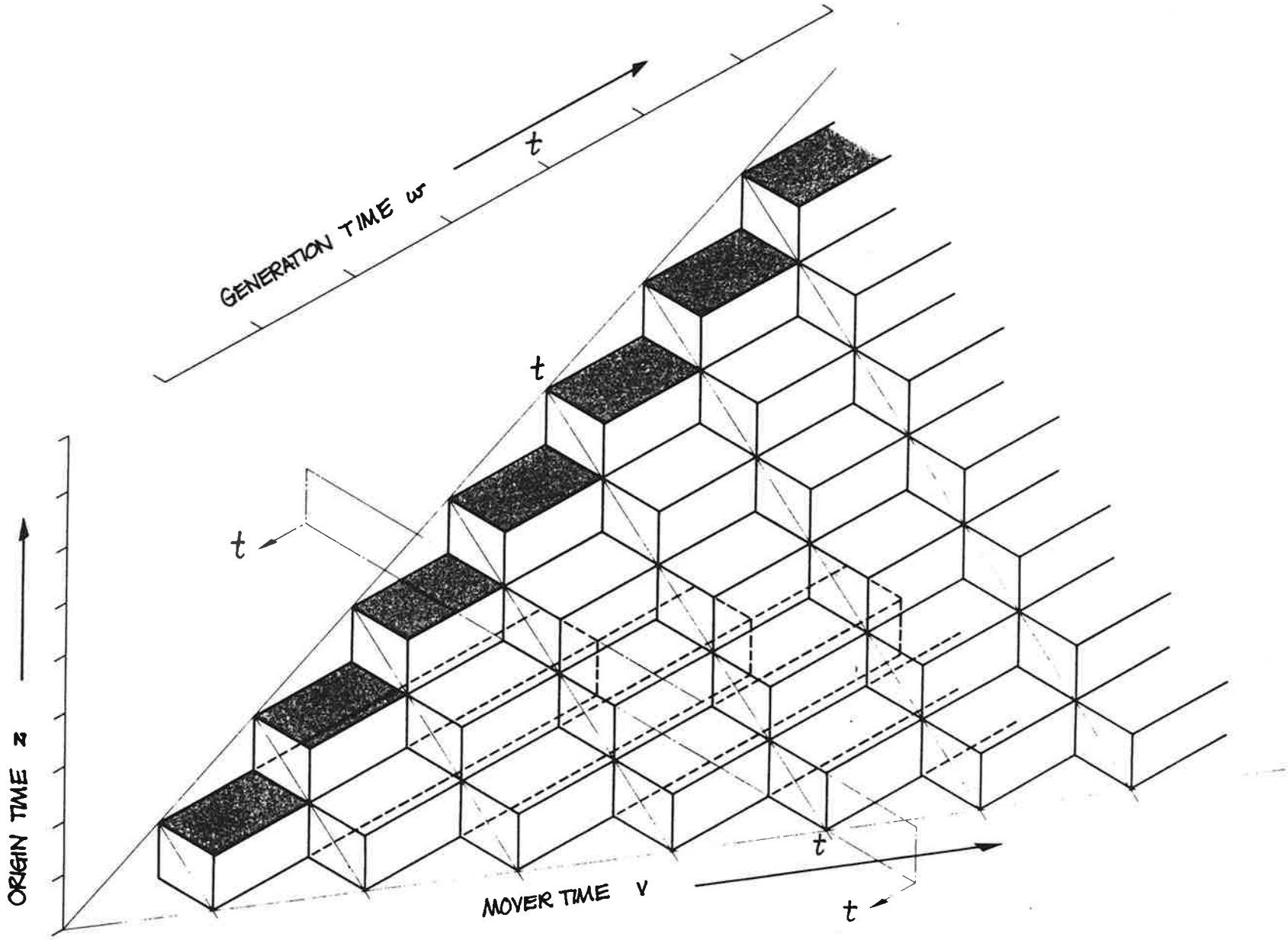



Figure 3.2: The Generation of Mover Streams.

The rest of this section is fairly straightforward, involving substition of equation (3.25) into equation (3.8) and making the appropriate summations. Note that the distinction between movers who are associated with activity which has already generated its complete sequence of new change, and movers whose activity base is still working itself out, can be clearly seen from Figure 3.2. If a slice is taken at time $\tau$ in Figure 3.2, then the logic for this distinction becomes visually apparent. First for service movers associated with activity originating prior to time period [t-T:t-T-1]

$$\sum_{z=0}^{t-T-1} \sum_{w=z}^{z+T} \underline{s}^{m}(t,w,z) = \sum_{z=0}^{t-T-1} \Delta\underline{b}(z) \sum_{v=t-T}^{t} \underline{\alpha}(t,v) \prod_{\tau=v}^{t} \underline{\tilde{A}}(\tau,z)\underline{\tilde{B}}(\tau,z), \qquad (3.26)$$

and for activity originating between t-T and t-1

$$\sum_{z=t-T}^{t-1} \sum_{w=z}^{t-1} \underline{s}^{m}(t,w,z) = \sum_{z=t-T}^{t-1} \Delta\underline{b}(z) \sum_{v=z+1}^{t} \underline{\alpha}(t,v) \prod_{\tau=v}^{t} \underline{\tilde{A}}(\tau,z)\underline{\tilde{B}}(\tau,z). \qquad (3.27)$$

For completeness, analogous equations exist for population movers and these are given as

$$\sum_{z=0}^{t-T-1} \sum_{w=z}^{z+T} \underline{p}^{m}(t,w,z) = \sum_{z=0}^{t-T-1} \Delta\underline{b}(z) \sum_{v=t-T}^{t} \underline{\alpha}(t,v) \prod_{\tau=v}^{t-1} \underline{\tilde{A}}(\tau,z)\underline{\tilde{B}}(\tau,z)\underline{\tilde{A}}(t,z), \qquad (3.28)$$

$$\sum_{z=t-T}^{t-1} \sum_{w=z}^{t-1} \underline{p}^{m}(t,w,z) = \sum_{z=t-T}^{t-1} \Delta\underline{b}(z) \sum_{v=z+1}^{t} \underline{\alpha}(t,v) \prod_{\tau=v}^{t-1} \underline{\tilde{A}}(\tau,z)\underline{\tilde{B}}(\tau,z)\underline{\tilde{A}}(t,z). \qquad (3.29)$$

Equations (3.26) to (3.30) determine the movers in any time period [t:t-1] and these are necessary before the stayers can be calculated. The process whichrelates stayers to movers is discussed in the next section.

THE COMPUTATION OF STAYERS.

The processes described so far are relatively easy to comprehend because in the case of both new change and movers, the sequence of change begins from a fixed distribution: basic employment. Thus these processes show how new patterns are built up on top of what already exists but the relationship between the process of redistributing existing patterns and calculating what remains of the existing - the stayers - is extraordinarily complex. In essence, just as the proportions $\underline{\alpha}(t,z)$ of the previous configuration of activities in the system form the potential movers, the residue $[\underline{I}-\underline{\alpha}(t,z)]$ form the stayers. However, the previous configuration of the system depends upon the movers and stayers at the previous time period, and this recurrence can be traced back indefinitely, in theory to the beginning of time (which is notionally taken as t=0 in this model).

To gauge the complexity of the process of finding an explicit equation to calculate stayers, consider the movers $\underline{s}^m(t,w,z)$ in any time t. The stayers $\underline{s}^s(t,w,z)$ can be calculated as a residual proportion of the movers and stayers who form the configuration of services at t-1 but this is only pushing the problem one stage further back. The stayers have to be calculated now at t-1 and so on. An equation for stayers can be derived by successive substitution back to the beginning of time but this is unnecessary. It will suffice to state the general recurrence relation for stayers and the initial conditions for the mover-stayer process.

To handle these various migration streams, it is necessary to define

a new variable $\underline{s}(\tau,w,z)$ which is a vector giving the location of services at time $\tau$, generated at w and originating at z. It is clear that this state variable is based on movers and stayers in the associated time period, and thus

$$\underline{s}(\tau,w,z) = \underline{s}^m(\tau,w,z) + \underline{s}^s(\tau,w,z), \quad \tau>w\geq z, \qquad \text{and}$$

$$\underline{s}(\tau,w,z) = \underline{s}(t,t), \quad \tau=w=z.$$

It is possible to take a particular component of service activity originating from $\Delta\underline{b}(t)$, say $\underline{s}(t,t)$,and trace the change in its configuration through time. This process can be traced diagrammatically in Figure 3.3 where the change in the original pattern $\underline{s}(\tau,t)$, $t\leq\tau\leq t+T$ is demonstrated. In the case of $\underline{s}(\tau,t)$, this activity is first generated at time t using equation (3.12) which can be rewritten as

$$\underline{s}(t,t,t) = \Delta\underline{b}(t)\underline{A}(t,t)\underline{B}(t,t). \tag{3.30}$$

In time period [t+1:t], the movers are given by equation (3.22), repeated here for convenience

$$\underline{s}^m(t+1,t,t) = \Delta\underline{b}(t)\underline{\alpha}(t+1,t+1)\underline{\tilde{A}}(t+1,t)\underline{\tilde{B}}(t+1,t) \qquad [(3.22)]$$

Stayers are calculated as a residual from equation (3.30)

$$\underline{s}^s(t+1,t,t) = \underline{s}(t,t,t)[\underline{I}-\underline{\alpha}(t+1,t+1)]. \tag{3.31}$$

A new configuration of services is then calculated by summing equations (3.22) and (3.31)

$$\underline{s}(t+1,t,t) = \underline{s}^m(t+1,t,t) + \underline{s}(t,t,t)\,[\underline{I}-\underline{\alpha}(t+1,t+1)] \tag{3.32}$$

Continuing in this fashion to time period [t+T:t+T-1] and concentrating upon the equation for stayers leads to

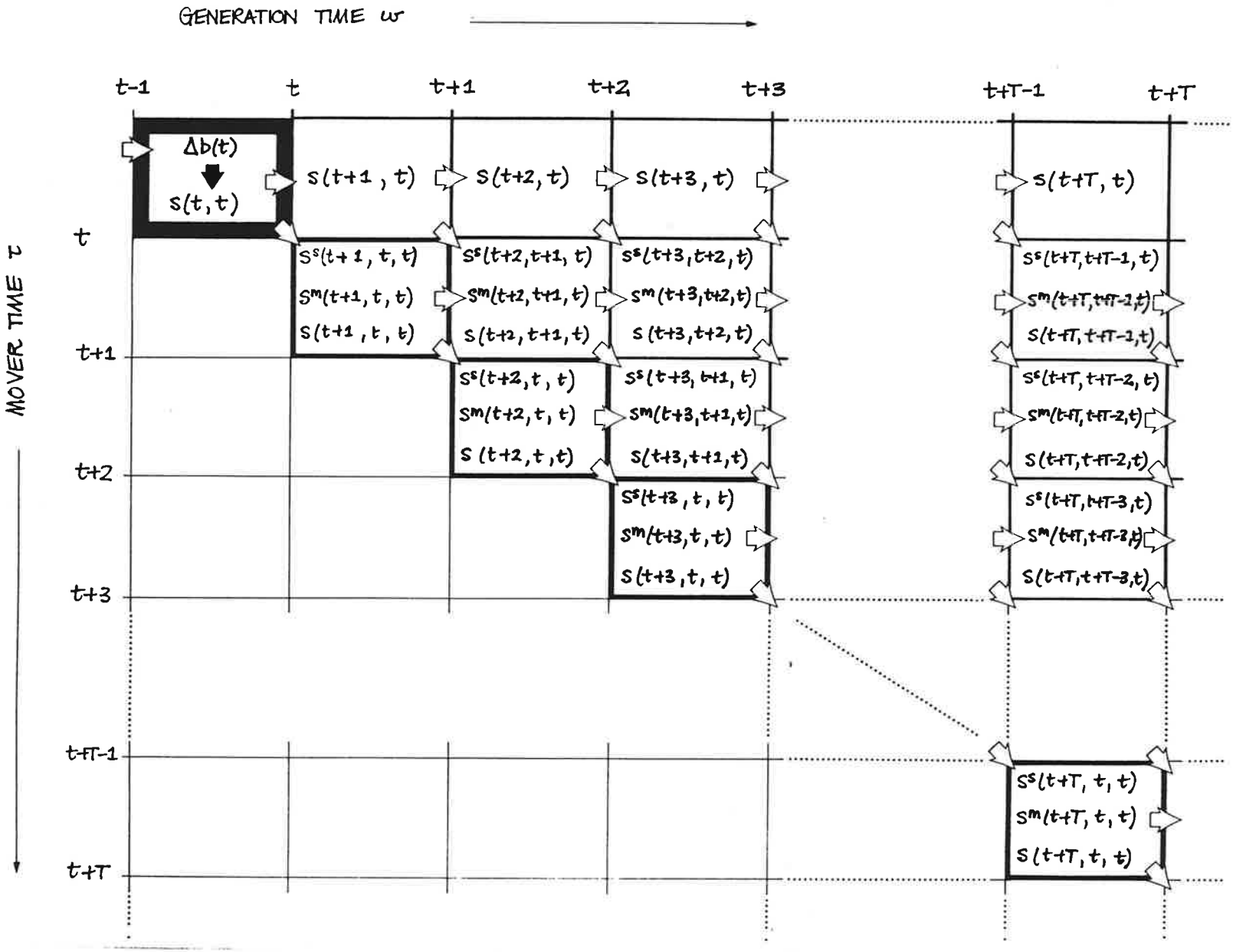



Figure 3.3: Changes to an Initial Pattern of Activity through Successive Redistribution.

$$\underline{s}^s(t+T,t,t) = \underline{s}(t+T-1,t,t)[\underline{I}-\underline{\alpha}(t+T,t+T)],$$
$$= [\underline{s}^m(t+T-1,t,t)+\underline{s}^s(t+T-1,t,t)][\underline{I}-\underline{\alpha}(t+T,t+T)], \quad (3.33)$$

which is the appropriate recurrence relation.

More generally for any time $\tau, \tau>z$, the amount of activity generated at w, originating at z which is stable is given by

$$\underline{s}^s(\tau,w,z) = [\underline{s}^m(\tau-1,w,z) + \underline{s}^s(\tau-1,w,z)][\underline{I}-\underline{\alpha}(\tau,\tau-w+z)], \quad (3.34)$$

with the initial conditions given by

$$\underline{s}^m(w+1,w,z) = \Delta\underline{b}(z)\underline{\alpha}(w+1,z+1) \prod_{\tau=z+1}^{w+1} \underline{\tilde{A}}(\tau,z)\underline{\tilde{B}}(\tau,z), \quad (3.35)$$

and

$$\underline{s}^s(w+1,w,z) = \Delta\underline{b}(z)[\underline{I}-\underline{\alpha}(w+1,z+1)] \prod_{\tau=z}^{w} \underline{A}(\tau,z)\underline{B}(\tau,z). \quad (3.36)$$

Note that the mover pool ratio in equation (3.34) has the range $\tau$-T to $\tau$ due to the fact that the generation process indexed by w is only significant in the interval $z \leq w \leq z+T$.

The final stage in explicitly computing the stayers is to sum the stayers over all origin times z and generation times w, to find the stayers in any time period [t:t-1]. As previously, it is necessary to distinguish between stayers whose activity base has been generated completely and stayers whose base has only been partly generated. The equation for total stayers is thus

$$\underline{s}^s(t) = \sum_{z=0}^{t-T-1} \sum_{w=z}^{z+T} \underline{s}^s(t,w,z) + \sum_{z=t-T}^{t-1} \sum_{w=z}^{t-1} \underline{s}^s(t,w,z). \quad (3.37)$$

By substituting equation (3.34) into (3.37) and rearranging summation indices, total stayers can be calculated from

$$\underline{s}^s(t) = \sum_{\tau=0}^{T} \{ \sum_{z=0}^{t-T-1} [\underline{s}^m(t-1,z+\tau,z) + \underline{s}^s(t-1,z+\tau,z)]\}[\underline{I}-\underline{\alpha}(t,t-\tau)]$$

$$+ \sum_{z=t-T}^{t-1} \sum_{w=z}^{t-1} [\underline{s}^m(t-1,w,z) + \underline{s}^s(t-1,w,z)][\underline{I}-\underline{\alpha}(t,t-w+z)]. \quad (3.38)$$

Analogous equations exist for population stayers with $\underline{p}^m(\tau,w,z)$ replacing $\underline{s}^m(\tau,w,z)$ in equations (3.37) and (3.38). The initial conditions for population are slightly different from equations (3.35) and (3.36) with the matrix product being taken over the range consistent with the process defined by equations (3.28) and (3.29). At this point in the development of the framework, the complete model can be assembled as a set of processes, and in the following section, an analysis of the properties of the model will be attempted.

BEHAVIOUR OF THE DYNAMIC MODEL.

Substituting the model equations for stayers, movers and new change into equations (3.5) and (3.6) does not lead to any new results but it does demonstrate the intricate nature of the framework and the varying richness and complexity of its parts. The equation for total employment (population has the same structure) is given here, and despite its cumbersome nature,it illustrates the various processes at work. Each term in the equation is presented on a separate line and these are organised from simple to complex. Then

$$\underline{e}(t) = \underline{b}(t-1) + \Delta\underline{b}(t)$$

$$+ \sum_{z=t-T}^{t} \Delta\underline{b}(z) \prod_{\tau=z}^{t} \underline{A}(\tau,z)\underline{B}(\tau,z)$$

$$+ \sum_{z=t-T}^{t-1} \Delta\underline{b}(z) \sum_{v=z+1}^{t} \underline{\alpha}(t,v) \prod_{\tau=v}^{t} \tilde{\underline{A}}(\tau,z)\tilde{\underline{B}}(\tau,z)$$

$$+ \sum_{z=0}^{t-T-1} \Delta\underline{b}(z) \sum_{v=t-T}^{t} \underline{\alpha}(t,v) \prod_{\tau=v}^{t} \tilde{\underline{A}}(\tau,z)\tilde{\underline{B}}(\tau,z)$$

$$+ \sum_{z=t-T}^{t-1} \sum_{w=z}^{t-1} [\underline{s}^m(t-1,w,z) + \underline{s}^s(t-1,w,z)][\underline{I}-\underline{\alpha}(t,t-w+z)]$$

$$+ \sum_{\tau=0}^{T} \{ \sum_{z=0}^{t-T-1} [\underline{s}^m(t-1,z+\tau,z) + \underline{s}^s(t-1,z+\tau,z)]\}[\underline{I}-\underline{\alpha}(t,t-\tau)] \tag{3.39}$$

The first line of equation (3.39) gives the exogenous basic employment which is input to the model, the second line gives the new change which is a simple function of the input in the previous T+1 time periods. The third and fourth lines show the movers who are a function of processes originating in the T+1 previous time periods, and involve *all* previous changes in the input. The fifth and sixth lines relate to stayers who like movers have the same dependence upon processes in the previous T+1 periods but are a function of the movers and stayers associated with all previous time periods. In the case of both movers and stayers, activities associated with the redistribution of new change processes still working themselves out, are stated first, on the third and fifth lines.

It is now possible to examine briefly the equilibrium behaviour of this model, before looking at the data problem involved in its construction. If the input $\Delta\underline{b}(t)$ were to cease, then t+T time units later, the last new change would be generated, and from then on change would be solely due to movement of the existing stock. Only if the mover ratio matrix $\underline{\alpha}(t,z)$ were to become null would the system approach a static position, and this would be of little interest. However, it is useful to examine the effect of stable movement patterns on the system. It seems intuitively likely that if the input ceases completely, and the pattern of redistribution becomes stable, then given enough time, some kind of steady state will be reached in which the system simply reproduces itself. Even if the re-

distribution matrices $\underline{\tilde{A}}(\tau,z)$ and $\underline{\tilde{B}}(\tau,z)$ do not become constant over $\tau$ and z, it seems likely that the earlier configuration of activities in the system will make a weaker and weaker contribution to the present pattern as time increases.

To illustrate these effects, consider the stayers associated with the stream $\underline{s}(t,t,t)$. Because of the additive nature of the mover-stayer process in the model, the analysis of any particular stream can be generalised to all streams, and thus it is appropriate to look at changes in $\underline{s}(t,t,t)$. Stayers associated with this stream $\underline{s}^s(t+T,t,t)$ at any time t+T are given by the general recurrence relation for stayers presented in equation (3.34). In this case

$$\underline{s}^s(t+T,t,t) = [\underline{s}^m(t+T-1,t,t) + \underline{s}^s(t+T-1,t,t)][\underline{I}-\underline{\alpha}(t+T,t+T)] \quad (3.40)$$

where the time index T is now being used as a general index for the increase in time, not solely as the fixed time of the generation process. Explicit recurrence on equation (3.40) back to time t leads to

$$\underline{s}^s(t+T,t,t) = \sum_{\tau=t+1}^{t+T-1} \{\underline{s}^m(\tau,t,t) \prod_{z=\tau}^{t+T-1} [\underline{I}-\underline{\alpha}(z+1,z+1)]\} + \underline{s}(t,t,t) \prod_{z=t+1}^{t+T} [\underline{I}-\underline{\alpha}(z+1,z+1)]. \quad (3.41)$$

The positive elements of the diagonal matrix $[\underline{I}-\underline{\alpha}(z,z)]$, $t+1 \leq z \leq t+T$, must fall in the range $0<\alpha_{ii}(z,z)\leq 1$ for the process to be able to generate movers, and thus it is clear that the last product term in equation (3.41) approaches the null matrix $\underline{0}$ at $T\to\infty$. Formally,

$$\lim_{T\to\infty} \prod_{z=t+1}^{t+T} [\underline{I}-\underline{\alpha}(z,z)] = \underline{0}, \qquad 0\leq\alpha_{ii}(z,z)\leq 1.$$

This is an explicit demonstration that the effect of the original pattern of activities $\underline{s}(t,t,t)$ dies away as the pattern is successively reordered through time. It also illustrates the fact that not only the original pattern, but all changes to the original pattern in the form of movers, die away as the time from the move increases. The same kind of limiting argument can thus be applied to the first set of product terms in equation (3.41).

To give some substance to this effect, assume that $\alpha_{ii}(z,z)$ is constant for all zones i and time z. Then each term on the main diagonal of the above matrix product has the form $(1-\alpha)^T$ where T is the number of time units from the origin of activity at t. Clearly $(1-\alpha)^T \to 0$ as $T \to \infty$, and it is worth noting how quickly some limit is reached for various 'typical' values of $\alpha$ and T. A reasonable limit for $(1-\alpha)^T$ to come within, is 0.01 which would imply that only 1 percent of the original pattern would contribute to the total pattern at time t+T; to all intents and purposes, it might be argued that the effect of the original pattern has then become insignificant. Realistic values of $\alpha$ are difficult to measure from population census data, for movers within the same spatial unit are missed, but it appears that for relatively prosperous communities $\alpha$ might be as large as 0.2 per annum: that is, 20 percent of activity makes a move each year. It would take 21 years for only 1 percent of the original pattern to remain in this case, and after 10 years, some 11 percent would remain.

However, a typical value for $\alpha$ is more like 0.05 per annum in South East England, and in this case, it would take 89 years for the pattern to become insignificant in terms of the 1 percent level. In some

senses, this type of mover behaviour is highly unrealistic for it presupposes that all activity is potentially movable. This is clearly not the case for some activity, for example, certain population groups and service sectors are by their nature inert for long periods. This could be handled through the mover pool mechanism in a disaggregated version of the above model, but it may be possible to treat it in the manner used in the semi-Markov model where the distribution matrices themselves are partitioned into movers and stayers (Bartholomew, 1982; Curry and MacKinnon, 1975). This is a matter for further research but it appears a promising line of inquiry.

A much simplified version of the model in equation (3.39) has already been built for the Reading subregion (Batty, 1976). The full model contained in this chapter however, has enormous computer storage requirements, and thus simplification would be necessary if it were to be made operational. To demonstrate the size of the problem consider the number of matrices which have to be distinguished, and held in store during a simulation. Table 3.1 gives a list of the key variables (excluding explicit intermediate variables) which are essential for the operation of the model over a simulation of T' years (T'-1 time periods). It is clear from this table that the critical storage requirements relate to the matrices $\underline{A}(\tau,z)$, $\underline{B}(\tau,z)$, $\underline{\tilde{A}}(\tau,z)$ and $\underline{\tilde{B}}(\tau,z)$. The storage requirements presented in this table could be reduced if information is output as soon as it is computed, and if the storage space for this information is continually reused as soon as the information is no longer needed. Here it is assumed that such space is used for the intermediate variables involved in the computation. A typical problem might involve, say,a total life of the generation processes of T=10, a simulation period of T'=15 and a variable number

Table 3.1: Dimensions of the Dynamic Model.

| Variables | Dimensions | | | |
|---|---|---|---|---|
| | Zones | Origin Time | Generation Time | Simulation Time |
| $\underline{b}(t)$ | N | | | T' |
| $\underline{s}(t)$ | N | | | T' |
| $\underline{e}(t)$ | N | | | T' |
| $\underline{p}(t)$ | N | | | T' |
| $\Delta\underline{b}(t)$ | N | T' | T+1 | T' |
| $\underline{s}(\tau,z)$ | N | T' | T+1 | |
| $\underline{p}(\tau,z)$ | N | T' | T+1 | |
| $\underline{s}^m(t)$ | N | | | T' |
| $\underline{s}^m(t,w,z)$ | N | T' | T+1 | T' |
| $\underline{s}^s(t)$ | N | | | T' |
| $\underline{s}^s(t,w,z)$ | N | T' | T+1 | T' |
| $\underline{p}^m(t)$ | N | | | T' |
| $\underline{p}^m(t,w,z)$ | N | T' | T+1 | T' |
| $\underline{p}^s(t)$ | N | | | T' |
| $\underline{p}^s(t,w,z)$ | N | T' | T+1 | T' |
| $\underline{A}(\tau,z)$ | $N^2$ | T' | T+1 | |
| $\underline{B}(\tau,z)$ | $N^2$ | T' | T+1 | |
| $\underline{\tilde{A}}(\tau,z)$ | $N^2$ | T' | T+1 | |
| $\underline{\tilde{B}}(\tau,z)$ | $N^2$ | T' | T+1 | |
| $\underline{\alpha}(t,z)$ | N2 | T' | | T' |
| $\underline{I}$ | $N^2$ | | | |

Total Dimension $= N\{T'\,[9+2(T+1) + 4T'(T+1) + 4N(T+1) + NT'] + N\}$

of zones N. In the Reading example, N=18 and this requires oves 473K of computer store. If the number of zones is increased to 50, the store required increases by a factor of $5\frac{1}{2}$ to over 2733K.

These are quite reasonable problems, for the degree of disaggregation of the time dimensions is commensurate with the spatial disaggregation. But it is clear that some simplification is necessary for feasible computation and this can be achieved in several ways. It may be possible to disregard previous service and population patterns if the simulation period and mover ratios are large enough for initial patterns to become insignificant. However this is unlikely in the context of the given example which has only 15 time points in the simulation. A more acceptable way, and one which was used in the Reading model, is to aggregate the distribution matrices over the generation times, that is, to form new matrices which have the form $\underline{A}(\tau,z) = \underline{T}(\tau)\Lambda(z)$ where the matrices $\underline{T}(\tau)$ and $\underline{\Lambda}(z)$ are as defined previously in relation to equations (3.19) and (3.20). Such a simplification reduces the store required for the 18 zone example by 38 percent, and for the 50 zone example by 51 percent. Other aggregations can be made but all will depend upon the specific application and the hypotheses concerning change processes which are testable from any given data base. Examples of these types of aggregation will be developed in the next chapter.

Now that the structure of the model and its generation processes have been outlined, it is necessary to turn to the processes of locating activities through the distribution matrices, for these too can have a dynamic form. To derive appropriate forms for matrices such as $\underline{A}(\tau,z)$, there already exist consistent frameworks based on generating

the most likely forms of statistical distribution (Wilson, 1974) and on other concepts such as the maximisation of economic utility (Beckmann, 1974). In this chapter, Wilson's (1970) information-theoretic framework will be generalised to a dynamic form, and various forms of lagged relationship will be derived. The approach builds on previous work by Batty and March (1978) which developed in parallel with the work of Snickars and Weibull (1977) and Webber (1979).

DYNAMIC INTERACTION MODELS: DERIVATION BY INFORMATION-MINIMISING.

In developing a method for deriving the form of the matrices such as $\underline{A}(\tau,z)$, it is necessary to distinguish between the scale and distribution effects. It is assumed that the scale effects are determined exogenously in accord with the hypotheses concerning generation, therefore the framework of this section is useful only in the derivation of distribution models. It is convenient if the reader has in mind the separability of these effects according to equations (3.19) and (3.20); in this case, the present section is concerned with estimating the matrix components such as $\underline{T}(\tau,z)$ of $A(\tau,z)$ although there are other ways of incorporating scale into $\underline{A}(\tau,z)$ than that given in equation (3.19). In the following argument, it will also be assumed that the elements of $\underline{T}(\tau,z)$ can be represented in probability form. Ways of expressing these probabilities in the distribution matrices are relatively simple and will be briefly mentioned later, and in greater detail in the next two chapters.

Just as the generation of new activity depends on past activity, so the distribution of this activity to zones is likely to depend on past distribution patterns. A consistent way of deriving models which adopt

this assumption is to use a framework in which a least prejudiced estimate of a new distribution is made in terms of some previous distribution and the change in that distribution's characteristics over the time period of interest. The characteristics of any distribution are referred to as information and a framework must be used in which this information is encoded into the new 'predicted' distribution through constraints on its form. Given such a framework, it would then be possible to use it recursively to continually 'update' the distribution in terms of new information. Clearly, information would be exogenous to the model, in the sense that it represents the force governing changes in distribution.

A static version of this framework in which a least prejudiced estimate of the form of some system is derived subject to information about the *absolute* structure of the system, is the entropy-maximising method originally due to Shannon (1948), popularised by Jaynes (1957) and extensively applied in urban modelling by Wilson (1970). However, there are philosophical difficulties concerning the measurement of absolute information and increasingly, the entropy-maximising method is being challenged on this question. The argument has been elaborated in diverse ways, and it does appear that a more general framework is one which uses information *relative* to some base, whether that base be a previous point in time or not. Indeed, even the use of Shannon's (1948) equation in communications engineering has been based on measuring information differences, rather than absolute information.

Relative information can be used to generate least prejudiced models by minimising a function relating the predicted 'posterior' distribution

to the known 'prior' distribution, subject to the information change characterising the prior distribution. Rather than maximising an entropy function relating to uncertainty, a function of information-gain is minimised. There are many such information functions (Reyni, 1960) but the function adopted here is well-known in the social sciences, and has been widely used by Kullback (1959) and Theil (1972). This framework for deriving least prejudiced models originates with the work of Hobson and Cheng (1973) but an early version can be found in Perez (1967). In urban modelling, these ideas continue those found in the papers by Batty and March (1976; 1978) and by Snickars and Weibull (1976) and Webber's (1979) book.

More formally, the idea is to minimise an information function $I(\tau,w)$, $\tau>w$, subject to the expected values of information $x^k(\tau,w)$, where the index $\tau$ relates to the present time, and w relates to some previous time. Note that w is not being used here as a generation time index but is being used as a general index to denote the existence of a distribution at a previous point in time. There are K+1 pieces of information $x^k(\tau,w)$, $k=0,1,\ldots,K$, and these may also be indexed in terms of zones i, j or the interaction between zones ij. In the following discussion, the probabilities of interaction between any zone i and any zone j at times $\tau$ and w will be denoted $p_{ij}(\tau)$ and $p_{ij}(w)$, and information change pertaining to this spatial interaction is given as $x^k_{ij}(\tau,w)$. The difference $(\tau-w)$ gives the temporal order of the process: a first order process $\tau=t$, $w=t-1$, will be illustrated below and then the concept will be generalised.

The first information function $I(t,t-1)$ is defined, following Hobson

and Cheng (1973), as

$$I(t,t-1) = \sum_{ij}\sum p_{ij}(t)\ell n \frac{p_{ij}(t)}{p_{ij}(t-1)} \quad . \tag{3.42}$$

To derive a model for the distribution $\{p_{ij}(t)\}$, the information gain in equation (3.24) must be minimised subject to a set of constraints on the expected values of this distribution. There are K+1 such constraints which pertain to new characteristics of the posterior distribution at time t. These are written as

$$\sum_{ij}\sum p_{ij}(t)f^k[x_{ij}^k(t,t-1)] = \langle X^k(t,t-1)\rangle, \; k=0,1,\ldots,K, \tag{3.43}$$

where $f^k$ is some function of the information change $x_{ij}^k$ between time t-1 and t, and $\langle X^k(t,t-1)\rangle$ is the expected value of this function known exogenously. Note that $f^0$ is defined as the unit function and therefore $\langle X^0(t,t-1)\rangle = 1$, the normalisation constraint.

The process of minimising a function such as I(t,t-1) in equation (3.42) subject to constraints such as those in equation (3.43), is well known. A Lagrangean is formed and minimised with respect to the probability distribution $\{p_{ij}(t)\}$, and the undetermined multipliers which reflect the constraints. This yields a set of equations which can be solved simultaneously: usually the model is expressed as

$$p_{ij}(t) = Z^{-1}p_{ij}(t-1) \sum_{k>0} \exp\{-\mu^k f^k[x_{ij}^k(t,t-1)]\} \quad , \tag{3.44}$$

where Z is the normalisation or scaling constant defined as

$$Z = \sum_{ij}\sum p_{ij}(t-1) \sum_{k>0} \exp\{-\mu^k f^k[x_{ij}^k(t,t-1)]\}, \tag{3.45}$$

$\mu^k$, k=1,2,..,K are undetermined multipliers, and note that the normalisation multiplier $\mu^0$ which relates to Z, is determined by substitution (Batty and March, 1978).

Equation (3.44) defines a recurrence on $p_{ij}(t-1)$, therefore it is possible to express any probability $p_{ij}(t)$ as a function of any previous probability $p_{ij}(t-T)$ and the sequence of information change from t-T to t. Defining $L_{ij}(t,t-1)$ as the information operator which updates $p_{ij}(t-1)$ to $p_{ij}(t)$, then

$$L_{ij}(t,t-1) = Z^{-1} \sum_{k>0} \exp \{-\mu^k f^k [x^k_{ij}(t,t-1)]\} ,$$

and thus equation (3.44) can be written as

$$p_{ij}(t) = p_{ij}(t-1)L_{ij}(t,t-1). \tag{3.46}$$

Recurrence on equation (3.46) from the change in period [t-T:t-T-1] to [t:t-1] leads to

$$p_{ij}(t) = p_{ij}(t-T-1) \prod_{w=t-T-1}^{t} L_{ij}(w,w-1), \tag{3.47}$$

which gives a clear indication of the compounding effect of first order information change. Generalisation to n'th order information functions and models is quite obvious in that the lag in equation (3.47) would be (w,w-n) and the distribution at the starting point $\{p_{ij}(t-T-n)\}$. Some properties of these models are discussed in Batty and March (1978).

ALTERNATIVE LAG FUNCTIONS.

Although it is possible to derive a model with any length of lag by using the appropriate information function, models with a series of lags can only be consistently generated using information functions which relate to the series of information changes consistent with the lags. Thus to generate a model which predicts a probability $p_{ij}(t)$ as a function of the previous T probabilities and associated

information changes, an appropriate information function needs to be defined. Consider the composite information gain I(t) which is a sum of previous significant information gains from t-T to t.

$$I(t) = \sum_{w=t-T}^{t} I(t,w-1). \tag{3.48}$$

Other composite functions could be formed. A weighted summation of information in which each term I(t,w-1) was modified by its contribution to I(t), may be appropriate in some contexts, and the weights might be prespecified according to some other dynamic model, or may possibly be the subject of a calibration. The function in equation (3.48) has an interesting interpretation: it can be written out and rearranged as

$$I(t) = \sum_{ij}\sum_{w=t-T}^{t} p_{ij}(t)\ell n \frac{p_{ij}(t)}{p_{ij}(w-T)} ,$$

$$= \sum_{ij}\sum p_{ij}(t)[(T+1)\ell n\ p_{ij}(t) - \sum_{w=t-T}^{t} \ell n\ p_{ij}(w-1)], \tag{3.49}$$

and if the average of equation (3.49) is taken over T+1 time periods, the equation becomes

$$\bar{I}(t) = \frac{I(t)}{T+1} = -\frac{1}{T+1} \sum_{ij}\sum p_{ij}(t) \sum_{w=t-T}^{t} \ell n\ p_{ij}(w-1)$$

$$+ \sum_{ij}\sum p_{ij}(t)\ell n\ p_{ij}(t). \tag{3.50}$$

Equation (3.50) is the average of the expected information inaccuracies defined by Kerridge (1961) minus the uncertainty due to Shannon (1948).

The model derived by minimising I(t) or $\bar{I}(t)$ subject to constraints on the form of $\{p_{ij}(t)\}$, is one in which T'th order information changes are applied to the previous T probabilities. Then the constraints to which

$\{p_{ij}(t)\}$ is subject are

$$\sum_{ij}\sum p_{ij}(t)f^k[x^k_{ij}(t,w-1)]=\langle X^k(t,w-1)\rangle, \quad k=1,2,\ldots,K$$
$$w=t-T,t-T+1,\ldots,t, \qquad (3.51)$$

and

$$\sum_{ij}\sum p_{ij}(t) = \langle 1\rangle \quad . \qquad (3.52)$$

Note that for this composite model, the normalisation is only defined once, and thus it is stated separately. The derivation is thus subject to K(T+1)+1 constraints, and the model is then given by

$$p_{ij}(t) = Z^{-1} \prod_{w=t-T}^{t} p_{ij}(w-1) \sum_{k>0} \exp\{-\mu^k(w-1)f^k[x^k_{ij}(t,w-1)]\} \quad ,$$

$$= \prod_{w=t-T}^{t} p_{ij}(w-1)L_{ij}(t,w-1), \qquad (3.53)$$

The variables $L_{ij}(t,w-1)$ are appropriately defined and Z is the normalisation constant which ensures that equation (3.53) sums to unity over i and j. Note that this constant must be defined as part of one $L_{ij}(t,w-1)$, usually $L_{ij}(t,t-1)$ as in equation (3.46) but this choice is arbitrary.

Equation (3.53) like equation (3.44) is a recurrence relation and it is thus possible to express $p_{ij}(t)$ as a function of the series of significant information changes from the most previous probability distribution. Assuming that the previous T probabilities are significant in determining each of the T probabilities in equation (3.53), the equation for $p_{ij}(t)$ can be written as

$$p_{ij}(t) = p_{ij}(t-T-1) \prod_{\tau=t-T}^{t} L_{ij}(\tau,\tau-1) \prod_{w=\tau-T}^{\tau-1} p_{ij}(w-1)L_{ij}(\tau,w-1). \qquad (3.54)$$

When the time T=0, equation (3.54) simplifies to the first order model given previously as equation (3.46). One final point needs to be made before this type of logic is used to estimate the form of matrices such as $\underline{T}(\tau,z)$: the model in equation (3.53) or (3.54) is completely general and the parameters $\mu^k(w-1)$ will determine the importance of any information change in the model. If there has been no change between two time periods, then the information gain $I(t,w-1)$ will be zero, information change $x_{ij}^k(t,w-1)$ will be zero, and the parameters $\mu^k(w-1)$ will have to be calibrated for every application and thus the order of the model can be determined empirically. Figure 3.4 provides a diagrammatic illustration of the lagged structure of of such a model.

INFORMATION-MINIMISING IN THE DYNAMIC URBAN MODEL.

To conclude this discussion, an indication of the way in which dynamic information minimising can be used to generate forms for the model's distribution matrices, will be presented. Assume as previously that these matrices can be partitioned according to equations (3.19) and (3.20) which are repeated here for convenience

$$\underline{A}(\tau,z) = \underline{T}(\tau,z)\underline{\Lambda}(z), \quad \text{and} \qquad [(3.19)]$$

$$\underline{B}(\tau,z) = \underline{S}(\tau,z)\underline{\Gamma}(z) \ . \qquad [(3.20)]$$

The matrices $\underline{\tilde{A}}(\tau,z)$ and $\underline{\tilde{B}}(\tau,z)$ have a similar structure and the following argument can easily be generalised to these mover redistribution matrices. Then as the matrices $\underline{T}(\tau,z)$ and $\underline{S}(\tau,z)$ deal with distribution from one zone to another, their cells contain the proportion or probability that an activity originating in i will locate or interact with another at j. To preserve the activity generation processes, these probabilities

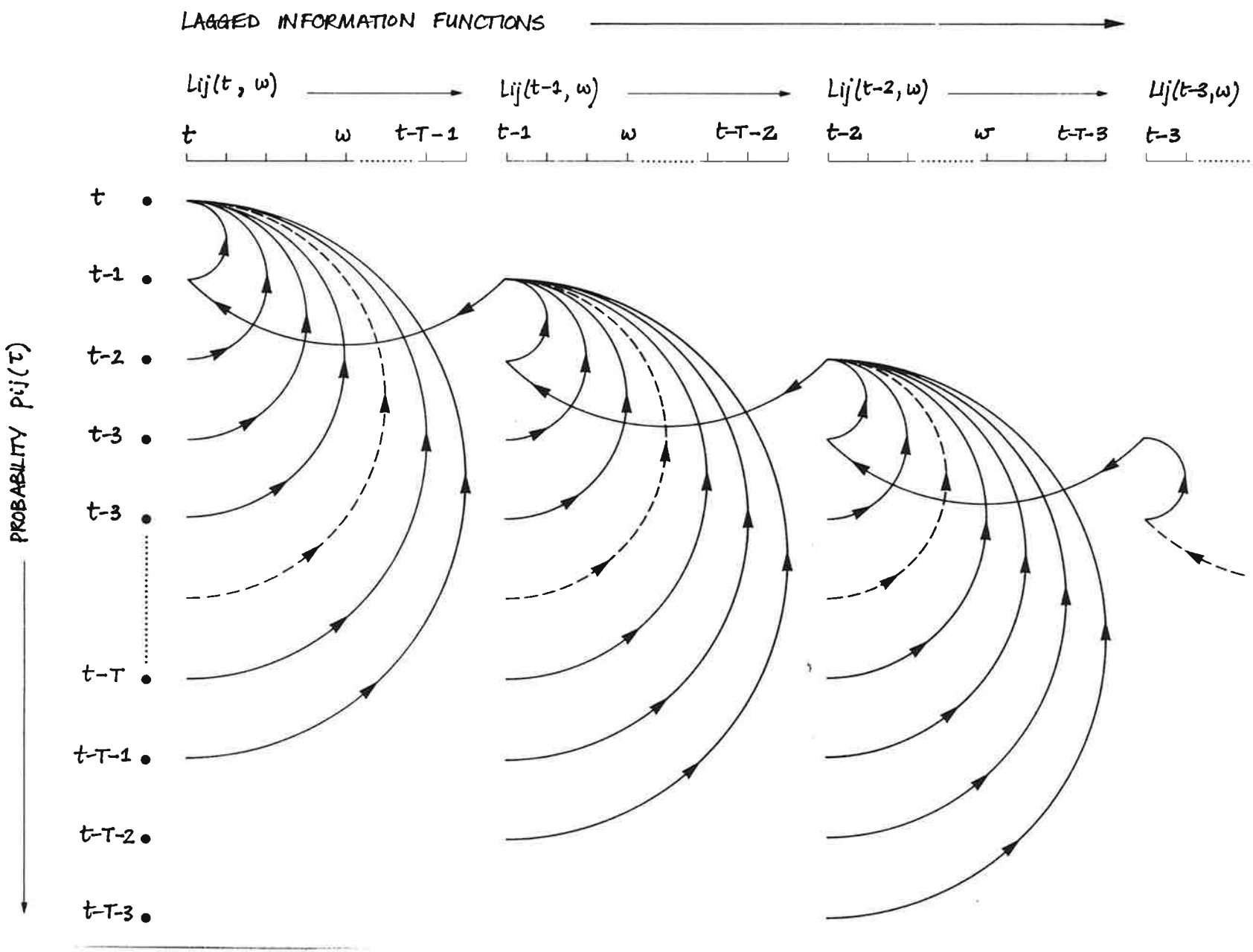



Figure 3.4: Lagged Structure of the Allocation Model.

are normalised to unity for each activity interaction emanating from zone i. In short, these matrices are singly stochastic and their elements can be easily derived from the probability interaction models presented above, if these probabilities are first normalised to unity by summing over j and scaling.

Consider a first order lag: using the above framework, it is possible to update matrix $\underline{T}(\tau-1,z)$ by applying a matrix $\underline{F}(\tau,\tau-1,z)$ whose elements are computed from the information change operator $L_{ij}(\tau,\tau-1)$. Note that the matrix change operator is notated according to the origin of the generation process. The matrix $\underline{F}(\tau,\tau-1,z)$ may not be computed explicitly, for the model may not be operated in matrix terms, but such matrices could easily be found if desired, and may contain useful information about the dynamics of the process (Batty and March, 1978). Assuming that the period of significant information change is T+1 time periods, and that $\underline{G}(\tau,\tau-1,z)$ represents the information change matrix operator appropriate to $\underline{S}(\tau,z)$, matrix recurrence equations analogous to equation (3.47) can be stated as

$$\underline{T}(t+T,z) = \underline{T}(t-1,z) \prod_{\tau=t}^{t+T} \underline{F}(\tau,\tau-1,z), \qquad \text{and} \qquad (3.55)$$

$$\underline{S}(t+T,z) = \underline{S}(t-1,z) \prod_{\tau=t}^{t+T} \underline{G}(\tau,\tau-1,z). \qquad (3.56)$$

Generalisation of equations (3.55) and (3.56) to n'th order lags, and to the redistribution matrices characterising movers and stayers, should be quite clear. The specific details of the elements on which these matrices are based, and how they are derived using information-minimising can be found in the next chapter.

The complete dynamic model has now been elaborated through its activity generation processes characterising new change, movers and stayers, and through the generalised lags in locational interactions which determine the distribution processes. It is not worthwhile substituting T'th order equations based on forms similar to equation (3.54) into the general model equation (3.39), for the resultant form would be long and cumbersome, and would add little of value. However it is worthwhile illustrating how a T'th order process of information change affecting spatial interaction combines with an activity generation process based on a life of T+1 time periods.

The simplest demonstration is based on the simplest process, that of generating new change. Using matrix equations for $\underline{T}(\tau,z)$ and $\underline{S}(\tau,z)$ with elements calculated from equations similar to (3.54), the amount of new service activity generated at any time t $\Delta^*\underline{s}(t)$, can be written as

$$\Delta^*\underline{s}(t) = \sum_{z=t-T}^{t} \Delta\underline{b}(z)\{ \prod_{\tau=z}^{t} \underline{\Lambda}(z)\underline{\Gamma}(z)$$

$$[\underline{T}(\tau-T-1,z) \prod_{w=\tau-T}^{\tau} \underline{F}(w,w-1,z) \prod_{v=w-T}^{w-1} \underline{T}(v-1,w)\underline{F}(w,v-1,z)]$$

$$[\underline{S}(\tau-T-1,z) \prod_{w=\tau-T}^{\tau} \underline{F}(w,w-1,z) \prod_{v=w-T}^{w-1} \underline{S}(v-1,w)\underline{G}(w,v-1,z)] \quad . \tag{3.57}$$

In equation (3.57), the first line contains the summation of new change generated in time t which results from the product of previous distributions of new change associated with the input back to time period [t-T:t-T-1]. The second and third lines deal with the T'th order distribution processes for population and services - the journey to

work and population demand for services - and these illustrate in themselves lags in the significance of information back T+1 time periods.

To summarise, the activity process, the order of the distribution process and the significance of past information are all associated with lags of a maximum of T+1 time periods. Of course, these maximum lags could vary, and be dependent on time themselves which would further complicate the process. Even more complicated expressions would result by making similar substitutions for the movers and stayers, but enough has been presented to give the reader a taste for the richness of the model. In the rest of this chapter, pseudo-dynamic models will be explicitly derived as a basis for their development in subsequent chapters.

## A CLOSED FORM FOR THE DYNAMIC URBAN MODEL.

The general definition of a pseudo-dynamic model presented in the introduction, suggested that such a model could be derived by aggregating a fully dynamic model with respect to time. As a first step in this process, it is necessary to express the dynamic model in closed form over a fixed time interval so that the input and influences of previous changes in time are well-defined. Indeed in operating the dynamic model itself, some degree of closure is necessary with respect to the simulation period, for it is unlikely that all previous influences on change in the simulation period are known. Thus approximation and an arbitrary starting position must be adopted as in the application of this type of model to the Reading region (Batty, 1976)

and all changes outside the simulation period must be known, assumed or ignored.

In applying the model to a closed interval of time, a number of fairly strong assumptions must be made. The time interval of the simulation is set from time period [t-T:t-T-1] to [t+T+1:t+T], and this interval is formed in the following way. From time period [t-T:t-T-1] to [t:t-1], there is input of activity $\Delta\underline{b}(z)$, $t-T\leq z\leq t$: prior to time t-T, the system is empty, and thus the period [t-T:t-T-1] constitutes the beginning of the world from the point of view of the model. Approximations to the past history of the system up to time t-T-1 would be required at the start of the process if the dynamic model were to be applied in this way, and the closed form prediction would simply be added to the past history to generate the present.

At any point in time $t-T+\tau$, $0\leq\tau\leq t$, the total exogenous activity input so far must be positive; that is, $\Sigma_{z=t-T}^{t-T+\tau}\Delta\underline{b}(z)>\underline{0}$, thus $\Delta\underline{b}(t-T)>\underline{0}$, although $\Delta\underline{b}(t-T+\tau)\gtreqless\underline{0}$, $\tau>0$, as long as the cumulative input is positive. The life of the generation process is T time units, thus by time t, the last new changes in population and services associated with the first input $\Delta\underline{b}(t-T)$ have been generated. In similar fashion, the last changes associated with $\Delta\underline{b}(t)$ are generated in the period [t+T:t+T-1], and after t+T, no new changes occur. In short, by t+T, enough time has elapsed for the input to have completely worked itself out.

Activity is able to first move one time period after it has been generated and thus the first movers and stayers associated with the

first input occur in [t-T+1:t-T]. It is assumed that all activity is able to move at least once in the closed interval, and as the last new change is generated in time period [t+T:t+T-1], this activity generates its first movers in [t+T+1:t+T]. Therefore the closed interval runs from the time of the first input to the time when the last generated change from the last input, makes its first move: this interval is from t-T-1 to t+T+1. It is also assumed that the mover ratio $\underline{\alpha}(\tau,z)>\underline{0}$, $t-T+1\leq z\leq t+1$ and $\underline{\alpha}(\tau,z) = \underline{0}$, $z>t+1$. Thus the last input in time period [t:t-1] begins its sequence of moves in [t+1:t] using the ratio $\underline{\alpha}(t+1,t+1)$ and generates its last move in [t+T+1:t+T]. The organisation of the closed interval from t-T-1 to t+T+1 is shown in Figure 3.5 and from this diagram, it is clear that the interval can be analysed in three subintervals. In the following section, closed form equations for the employment vector will be derived, and as previously, analogous equations for population immediately follow and are thus not stated. Equation (3.30) will be used as a basis for these derivations.

It is clear from Figure 3.5 that the closed interval can be divided into three periods. In the first period [t-T:t-T-1], there is only new change at the start of the process for no movers are yet possible. From [t-T+1:t-T] to [t:t-1] inputs occur, and movers and stayers are generated from previous activity. Finally in the interval [t+1:t] to [t+T+1:t+T], there is no input, only new change and movers originating from previous inputs. The three subintervals are characterised by exogenous change, exogenous and endogenous change, and endogenous change respectively. In terms of the first two subintervals, the system is being driven externally whereas in the last subinterval

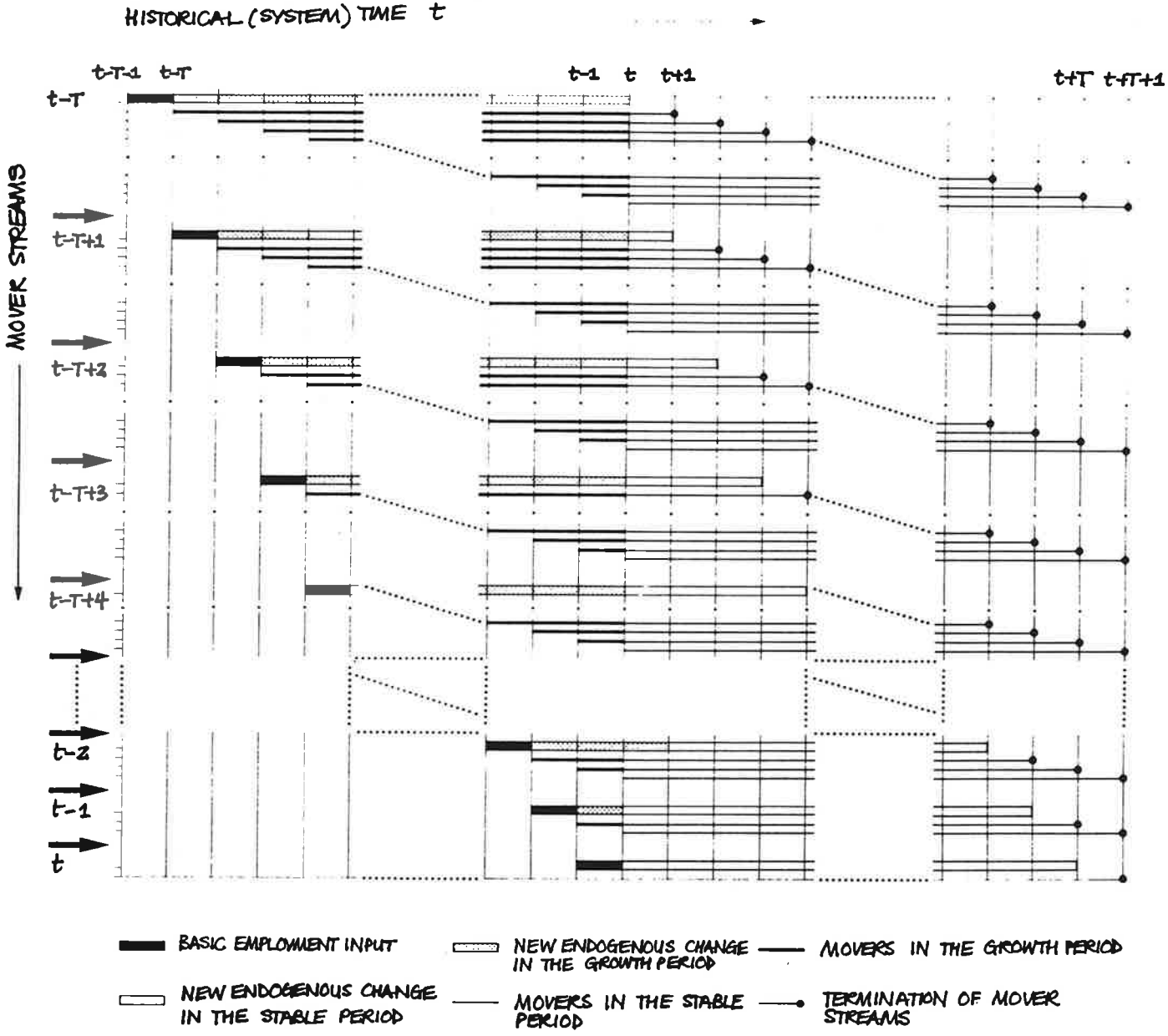



Figure 3.5: Temporal Organisation of the Dynamic Process in Closed Form.

it is approaching a kind of equilibrium as the momentum already established begins to lessen. It is necessary to distinguish between these subintervals because equation (3.30) simplifies in different ways for change within each of these periods. First for the period [t-T:t-T-1], employment is calculated from

$$\underline{e}(t-T) = \Delta\underline{b}(t-T) + \Delta\underline{b}(t-T)\underline{A}(t-T,t-T)\underline{B}(t-T,t-T). \qquad (3.58)$$

At t-T, there are no movers for these only begin in the following period [t-T+1:t-T], and thus from this period until time t, employment must be calculated using a more complicated expression. Then

$$\begin{aligned}\underline{e}(r) = & \sum_{z=t-T}^{r} \Delta\underline{b}(z) \\ & + \sum_{z=t-T}^{r} \Delta\underline{b}(z) \prod_{\tau=z}^{r} \underline{A}(\tau,z)\underline{B}(\tau,z) \\ & + \sum_{z=t-T}^{r-1} \Delta\underline{b}(z) \prod_{v=z+1}^{r} \underline{\alpha}(r,v) \prod_{\tau=v}^{r} \tilde{\underline{A}}(\tau,z)\tilde{\underline{B}}(\tau,z) \\ & + \sum_{z=t-T}^{r-1} \sum_{w=z}^{r-1} [\underline{s}^{m}(r-1,w,z) + \underline{s}^{s}(r-1,w,z)][\underline{I}-\underline{\alpha}(r,r-w+z)], \\ & \qquad t-T+1 \leq r \leq t, \qquad (3.59)\end{aligned}$$

Equation (3.59) is a simplification of equation (3.39) in that up to t, new change is still working itself out and there are no movers and stayers associated with past inputs which have completely generated their associated activity. Like equation (3.30), equation (3.59) is organised to demonstrate exogenous activity on its first line, endogenous on its second, third and fourth lines, namely new change, movers and stayers. Noting that $\Delta\underline{b}(r)=\underline{0}$, $r>t$, $\underline{\alpha}(\tau,r)=\underline{0}$, $r>t+1$, and

$\underline{A}(r,z)=\underline{0}$, $\underline{B}(r,z)=0$, $\tilde{\underline{A}}(r,z)=0$, $\tilde{\underline{B}}(r,z)=0$, $r>z+T$, equation (3.59) could be used to compute change in the subinterval $t+1\leq r\leq t+T+1$. However there is a clearer expression for change in this period.

Following the structure of equation (3.39) and using the definition $\underline{b}(t) = \Sigma_{z=t-T}^{t}\Delta\underline{b}(z)$, employment $\underline{e}(r)$, $t+1\leq r\leq t+T+1$, can be calculated from

$$
\begin{aligned}
\underline{e}(r) = \underline{b}(t) & \\
&+ \sum_{z=r-T}^{t} \Delta\underline{b}(z) \prod_{\tau=z}^{r} \underline{A}(\tau,z)\underline{B}(\tau,z) \\
&+ \sum_{z=r-T}^{t} \Delta\underline{b}(z) \sum_{v=z+1}^{t+1} \underline{\alpha}(r,v) \prod_{\tau=v}^{r} \tilde{\underline{A}}(\tau,z)\tilde{\underline{B}}(\tau,z) \\
&+ \sum_{z=t-T}^{r-T-1} \Delta\underline{b}(z) \sum_{v=r-T}^{t+1} \underline{\alpha}(r,v) \prod_{\tau=v}^{r} \tilde{\underline{A}}(\tau,z)\tilde{\underline{B}}(\tau,z) \\
&+ \sum_{z=r-T}^{\tau} \sum_{w=z+r-t-1}^{r-1} [\underline{s}^{m}(r-1,w,z)+\underline{s}^{s}(r-1,w,z)][\underline{I}-\underline{\alpha}(r,r-w+z)] \\
&+ \sum_{\tau=r-t-1}^{T} \sum_{z=t-T}^{r-T-1} [\underline{s}^{m}(r-1,z+\tau,z)+\underline{s}^{s}(r-1,z+\tau,z)][\underline{I}-\underline{\alpha}(r,r-\tau)] \\
&\quad \sum_{\tau=0}^{r-t-2} \sum_{z=t-T}^{t} \underline{s}^{s}(t+\tau+1,z+\tau,z), \quad t+1\leq r\leq t+T+1, \qquad (3.60)
\end{aligned}
$$

Equation (3.60) differs from equation (3.30) in a small but important way. The number of components which form the mover-stayer recurrence equations decreases as r increases, due to the fact that no new input occurs after the period [t:t-1] and no new mover sequences are begun after [t+1:t] . Those sequences which have begun previously finish and eventually in the period [t+T+1:t+T],the last movers are generated.

This decline in the number of new changes and movers is due to the termination of the input in [t:t-1] and the last of the mover sequences beginning in [t+1:t]. Clearly as the components of change get less, the number of stayer components increases to compensate, and the last line of equation (3.60) includes the stayers which are constant from r to the end of the simulation period. Note that these stayers only exist for $t+2 \leq r \leq t+T+1$. For r=t+1, the summation over $\tau$ is out of range and is assumed to be undefined. The stayers which make up this term, are those which represent the final pattern made up of movers and stayers to $t+\tau+1$ for each component, and these are unchanging after this time due to the termination of the associated mover sequence.

Another way of demonstrating this movement to a static situation at time t+T+1 can be illustrated by examining equation (3.60) at the end of the closed interval. Then at t+T, employment is given by

$$
\begin{aligned}
\underline{e}(t+T) = \; & \underline{b}(t) \\
& + \Delta\underline{b}(t) \prod_{\tau=t}^{t+T} \underline{A}(\tau,t)\underline{B}(\tau,t) \\
& + \Delta\underline{b}(t)\underline{\alpha}(t+T,t+1) \prod_{\tau=t+1}^{t+T} \tilde{\underline{A}}(\tau,t)\tilde{\underline{B}}(\tau,t) \\
& + \sum_{z=t-T}^{t-1} \Delta\underline{b}(z) \sum_{v=t}^{t+1} \underline{\alpha}(t+T,v) \prod_{\tau=v}^{t+T} \tilde{\underline{A}}(\tau,z)\tilde{\underline{B}}(\tau,z) \\
& + [\underline{s}^{m}(t+T-1,t+T-1,t)+\underline{s}^{s}(t+T-1,t)][\underline{I}-\underline{\alpha}(t+T,t+1)] \\
& + \sum_{\tau=T-1}^{T} \{ \sum_{z=t-T}^{t-1} [\underline{s}^{m}(t+T-1,z+\tau,z)+\underline{s}^{s}(t+T-1,z+\tau,z)]\}[\underline{I}-\underline{\alpha}(t+T,t+T-\tau)]
\end{aligned}
$$

$$+ \sum_{\tau=0}^{T-2} \sum_{z=t-T}^{t} \underline{s}^{s}(t+\tau+1,z+\tau,z). \qquad (3.61)$$

Equation (3.61)shows that at time t+T, the last new change associated with the input at time t, is generated (second line). In terms of new movers and stayers, there is only one component associated with the input at time t (third and fifth lines), and for movers and stayers whose generation process is fully worked out, there are two components for each input up to time t-1.

At time t+T+1, no new change is generated, and there is only one component for the movers and stayers originating from each input (from [t-T:t-T-1] to [t:t-1]). Then at t+T+1

$$\begin{aligned}\underline{e}(t+T+1) = \underline{b}(t) &+ \sum_{z=t-T}^{t} \Delta\underline{b}(z)\underline{\alpha}(t+T+1,t+1) \prod_{\tau=t+1}^{t+T+1} \tilde{\underline{A}}(\tau,z)\tilde{\underline{B}}(\tau,z) \\ &+ \sum_{z=t-T}^{t} [\underline{s}^{m}(t+T,z+T,z)+\underline{s}^{s}(t+T,z+T,z)][\underline{I}-\underline{\alpha}(t+T+1,t+1)] \\ &\sum_{\tau=0}^{T-1} \sum_{z=t-T}^{t} \underline{s}^{s}(t+\tau+1,z+\tau,z), \qquad (3.62)\end{aligned}$$

and it is clear that the process involves only the last components of change associated with the movers. Readers can also check this number of components of change from Figure 3.5 where the gradual movement towards the static position is traced diagrammatically. At time t+T+2, there are no movers, only stayers and the fixed input, thus $\underline{e}(t+T+2)=\underline{e}(t+T+1)$, and the process has reached a static equilibrium which defines the end of the simulation period. This closed form model can now be used as a basis for temporal aggregation.

THE DERIVATION OF PSEUDO-DYNAMIC URBAN MODELS.

In the fully dynamic model, there are two significant time streams which might be aggregated to provide more macro forms. First there is the origin time z and second, there is current system time t: as generation time w is always associated with an origin time z, w and z both index the same process, and thus it is only necessary to treat one of these, say z. It is now possible to provide a technical definition of a pseudo-dynamic model with the fully dynamic model in mind: a pseudo-dynamic model, then, is a model with an explicit dynamic form characterised by two or more significantly different time streams, some of which are aggregated and treated statically, others of which remain in their basic form. Such a model has both static and dynamic elements, and it goes without saying that the temporal aggregation must be accomplished in a meaningful way. In this and subsequent chapters, pseudo-dynamic models are derived from dynamic ones in two stages: first by expression of the dynamic model in closed form and then by temporal aggregation, although this does not necessarily appear to be the only way to derive such models. It has been introduced here purely as a matter of convenience.

Even at this stage, there are several different pseudo-dynamic models which could be derived from the closed form in equations (3.58) to (3.60). For example, it is possible to aggregate the input activity from t-T to t and to assume that the process begins in the period [t:t-1]. The process would be subject to exactly the same previous assumptions concerning $\underline{\alpha}(\tau,z)$, and it is thus clear that only one sequence of movers

would be generated beginning at [t+1:t]. However a richer model would be one in which it was assumed that the input was aggregated prior to the start of the process in [t-T:t-T-1], and that the process was the same from then on. This would imply that the activity would generate T+1 series of moves rather than only one, although the first model could be derived from the second, and in both cases, the essential idea is that the T+1 distinct streams of activity associated with the T+1 origin times z are aggregated to one stream, and treated according to the closed form dynamic model.

In these applications, it is assumed that it is the origin times which are aggregated, and this has implications for the system time t. No longer can t take on the same significance for this dimension now becomes associated with a single input, and thus it might be interpreted as a kind of composite system time, or more likely model time. In the event, it represents an approximation to the real dynamics as in any process of aggregation. The fuller implications of this point will be spelt out in the next chapter.

The second of the above pseudo-dynamic models will be elaborated here. First, consider that the process of aggregating the input $\Delta\underline{b}(z)$ is outside the simulation period [t+T+1:t-T-1]. Then

$$\underline{b} = \sum_{z=0}^{t-T} \Delta\underline{b}(z), \qquad [\Delta\underline{b}(z)=0, z>t-T]$$

where z=0 represents some notional origin of the system. In fact, in applications of such models, $\underline{b}$ would be measured directly at t-T. Given these two assumptions concerning the input, the process is then identical to that outlined in equations (3.58) to (3.60).

However, certain simplifications arise due to these assumptions, and thus it is worthwhile seeing how these are reflected in the equations.

Furthermore, it is necessary to derive an explicit form for this pseudo-dynamic model which will represent the starting point for later chapters. As the origin time z has been aggregated to one period, then variables no longer need to be notated by z. The equation for the first subinterval of the simulation period for exogenous growth only, analogous to (3.58) is

$$\underline{e}(t-T) = \underline{b} + \underline{b}\,\underline{A}(t-T)\underline{B}(t-T). \tag{3.63}$$

In the subinterval [t:t-T], employment $\underline{e}(r)$ is calculated from

$$\begin{aligned}\underline{e}(r) = {} & \underline{b} \\ & + \underline{b} \prod_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau) \\ & + \underline{b} \sum_{v=t-T+1}^{r} \underline{\alpha}(r,v) \prod_{\tau=v}^{r} \tilde{\underline{A}}(\tau)\tilde{\underline{B}}(\tau) \\ & + \sum_{w=t-T}^{r-1} [\underline{s}^{m}(r-1,w)+\underline{s}^{s}(r-1,w)][\underline{I}-\underline{\alpha}(r,r-w+t-T)], \\ & \qquad\qquad\qquad\qquad\qquad\qquad t-T\leq r\leq t,\end{aligned} \tag{3.64}$$

and it is clear that the four lines of equation (3.64) are analogous to the lines of equation (3.59): note that in equation (3.64), each term can be derived from (3.59) by suppressing z.

In the final subinterval from t+1 to t+T+1, the new change in the second line of equation (3.64) disappears and the process begins to lose its momentum immediately. An equation analogous to (3.60) but missing the mover-stayer terms associated with inputs generating

new change, can be derived as

$$\underline{e}(r) = \underline{b} + \underline{b} \sum_{v=r-T}^{t+1} \underline{\alpha}(r,v) \prod_{\tau=v}^{r} \underline{\tilde{A}}(\tau)\underline{\tilde{B}}(\tau)$$

$$+ \sum_{\tau=r-t-1}^{T} [\underline{s}^{m}(r-1,t-T+\tau)+\underline{s}^{s}(r-1,t-T+\tau)][\underline{I}-\underline{\alpha}(r,r-\tau)]$$

$$+ \sum_{\tau=0}^{r-t-2} \underline{s}^{s}(t+\tau+1,t-T+\tau), \quad t+1 \leq r \leq t+T+1. \qquad (3.65)$$

In the final period [t+T+1:t+T], the last movers are generated and equation (3.65) simplifies to

$$\underline{e}(t+T+1) = \underline{b} + \underline{b}\, \underline{\alpha}(t+T+1,t+1) \prod_{\tau=t+1}^{t+T+1} \underline{\tilde{A}}(\tau)\underline{\tilde{B}}(\tau)$$

$$+ [\underline{s}^{m}(t+T,t)+\underline{s}^{s}(t+T,t)][\underline{I}-\underline{\alpha}(t+T+1,t+1)]$$

$$+ \sum_{\tau=0}^{T-1} \underline{s}^{s}(t+\tau+1,t-T+\tau). \qquad (3.66)$$

At t+T+2, there are no movers, and the system is effectively static. The pseudo-dynamic model in equations (3.64) to (3.65) is the one which will be used in subsequent chapters in this thesis. Other models are possible and will be explored in future research, but the model derived here is sufficiently general to form a basis for reinterpreting existing static models and designing better ones.

CONCLUSIONS.

It has taken some time to present the idea of a pseudo-dynamic model

but a thorough understanding of the dynamic processes from which such a model is derived, is an essential prerequisite in the development of this concept. Of particular importance in the study of pseudo-dynamic models is the way in which explicit processes dealing with new change, movers and stayers have to be reinterpreted when temporal aggregation occurs, and in subsequent chapters such processes will take on new roles involving the control of the pseudo-dynamic process. Indeed, it is difficult to know how equations such as those in (3.63) to (3.65) could have been derived without the full development of the dynamic model, and as will become clearer later,this pseudo-dynamic model can be used to derive new forms of static model; and it is suggested that this line of inquiry is as relevant a way of improving existing models as other schemes such as sectoral disaggregation.

In the next chapter, the pseudo-dynamic model in equations (3.63) to (3.65) will be used in deriving a typology of models, some of which are already known, some of which are new. Insights into what existing models emphasise and what they do not are obtained, and these lead on to new ways of modelling constraints which these models have to meet. The information-minimising framework outlined above will also be used to generate specific forms of dynamic interaction model, and a particular pseudo-dynamic model will then be developed and applied to the Reading subregion in Chapter 5. This model will be calibrated in a static way, but in Chapter 6, a new method of calibration based on the dynamic structure of the process will be presented. This method is analogous to a process of optimal control of a dynamic system, and it is here that the notion of a pseudo-dynamic model helps in re-interpreting existing methods. In fact, the calibration method is

considerably more efficient than existing practice, and the use of the method on existing models which have the minimal of pseudo-dynamic structure, has obvious advantages. In Chapter 7, the method is explored on an urban model of Peterborough, while later chapters explore the calibration process using yet another variant of the pseudo-dynamic model.

These developments are based on the notion that improvements to existing urban models must come through reinterpretation and extensions to such models, not just through radically new formulations. The models which will be presented despite their apparent theoretical simplicity, have the prime advantage that they can be applied to real situations, and involve urban activities which are of potential interest to land use planning. Despite the fact that there are continuing calls for the abandonment of such models due to their simplistic structure or seeming irrelevance (see for example, Lee, 1973), they continue to be built. The difficulties which plague the development of theory in social science, should be enough to convince even the most extreme, that improvements to present practice, are worthy of time and effort, even if only to demonstrate the limitations on the state of the art.

# CHAPTER 4.

## A TYPOLOGY OF PSEUDO-DYNAMIC MODEL FORMS.

The pseudo-dynamic model introduced in the last chapter is designed to simulate processes of urban change whose form is implicit, rather than explicit. There are many situations where the phenomena of interest can only be explained in a temporal sense but frequently the dynamic processes which determine the phenomena cannot be observed due to lack of information or due to their intrinsic nature. In these circumstances, there appear to be two possible approaches: to assume the dynamics of the process and to build an explicitly dynamic model which cannot be tested against any available data, or to build a static model with an implicit dynamic structure which is capable of testing against cross-sectional data. The first approach is the one adopted by researchers such as Forrester (1969), the second is the approach postulated here which involves the construction and application of pseudo-dynamic models.

Although the idea of a pseudo-dynamic model was explored at length in the last chapter, it is worthwhile summarising its structure before the formal elaboration of the model begins. A pseudo-dynamic model is defined as a model in which both static and dynamic elements exist.

In essence, the model is neither wholly static nor dynamic for it contains elements which pertain both to cross-sectional and time-series data. Such a model can be derived from a fully dynamic model which is characterised by two or more related dynamic processes, by aggregating a subset of these processes with respect to time. The resulting model thus contains at least one dynamic process and at least one static approximation to such a process. There are several types of process which might be treated in this manner. For example, it may be necessary to design a static model in which the static nature of the phenomena is grown 'artificially' to the cross-section in time. Such are the models which originate from Lowry's (1964) idea of the 'instant metropolis'. Models in which constraints on the phenomena are necessary, can often be interpreted in pseudo-dynamic terms. As activity is built up by a model, constraints on the amount, its location, the patterns of interaction generated by it and so on come into play. Usually in static models, such constraints are met by iterative solution but often, a dynamic interpretation of the rationale for such constraints is meaningful, and thus a pseudo-dynamic treatment is relevant.

The model derived in Chapter 3 was based on the distinction between processes of new change, and processes of redistributing existing activity. In short, exogenous changes originating outside the system led to changes within the system which involve allocating urban activity to space whereas changes within the system led to the reallocation of that same activity to space at some time after it had first originated. The amount of activity reallocated was referred to as mover activity and activity which was stable as stayer activity. The pseudo-dynamic model retains this distinction between new change, movers and stayers, and the idea of allocation and reallocation

associated with these processes is a useful means of modelling the effect of constraints in static models. For example, models in which solutions have to be reached iteratively due to the simultaneous relationships between activities in an urban system may have to be solved and then re-solved to ensure consistency. Constraints on location or on the pattern of interaction between activities might be met by initial allocation followed by successive reallocation, and thus the structure of the pseudo-dynamic model seems eminently suited to handling these procedures.

In this chapter, the pseudo-dynamic model presented in equations (3.63) to (3.66) will be presented and extended a little further. By varying certain elements which characterise the model's processes, it is possible to generate a family or typology of related models and this chapter will be concerned with elaborating this typology. Some interesting insights into new and existing model forms are gleaned from this discussion, and then in the next chapter, the idea of using these models to simulate the effect of locational constraints is explored. At this point, an application of one of the models derived is presented. The model is first stated in a computable form and then the use of the information-minimising framework to derive the allocation and reallocation submodels is described. The resulting cross-sectional model is calibrated to the Reading subregion and a brief comment on its performance is made. The static nature of the calibration is somewhat arbitrary and this leads on to the notion of dynamic calibration through the model's pseudo-dynamic structure, the subject of the later chapters.

THE FORM OF THE PSEUDO-DYNAMIC MODEL.

The time period over which the model operates is divided into three main

intervals, together with an initial input stage. It is assumed that the initial input is calculated prior to the operation of the model, and that there is only one input driving the whole process through the simulation period. In fact, the aggregation of the input from each time period into one total input is achieved by aggregating the inputs of the fully dynamic model, thus deriving the pseudo-dynamic form and a detailed discussion of this has already been given. It is also assumed that the new change generated by the input has a life of T time units, and that the life of each process which involves redistributing the new change also has a life of T units. Furthermore, it is assumed that at every period after a new change has occurred in the system, a new process of re-allocation involving movers begins,and ends T units later.

These assumptions imply that the simulation period can be divided into three main intervals: the time period [t-T:t-T-1] the beginning of the simulation when only new change is generated, the period [t:t-T] which is characterised by new change, movers and stayers, and the period [t+T+1:t] which is characterised by only movers and stayers. At time period [t+T+2:t+T+1], the system is in equilibrium for there is no new change and no movers, only stayers. The model can now be formally stated: note that in the following presentation only equations for employment are presented, for population equations immediately follow as demonstrated earlier. However, in the next chapter on the application of locational constraints, and on the development of the model for the Reading subregion, the population equations will be given. Notation conventions are as in Chapter 3, that is, bold lower case letters are 1 x N row vectors and bold capitals are N X N matrices, N being the number of employment-population locations (zones) in the system.

Existing definitions of variables will be restated when necessary.

The initial input of basic employment $\underline{b}$ is aggregated from changes in such employment $\Delta\underline{b}(z)$ from the beginning of the system's history z=0 which is purely notional, to time t-T. Then

$$\underline{b} = \sum_{z=0}^{t-T} \Delta\underline{b}(z) \ , \quad [\Delta\underline{b}(z) = 0, \ z>t-T].$$

New change generated from the input $\underline{b}$ in time period [r:r-1], r>t-T, is called $\Delta^*\underline{s}(r)$. Movers associated with new change generated in time period [w:w-1], and associated with [r:r-1] are $\underline{s}^m(r,w)$ and stayers $\underline{s}^s(r,w)$. The model can now be written in terms of the three time intervals defined above. Readers should note that although this discussion is self-contained on the formal level, it may be necessary to occasionally consult the last chapter to refresh the model's logic.

In period [t-T:t-T-1], total employment $\underline{e}(t-T)$ is given by

$$\underline{e}(t-T) = \underline{b} + \Delta^*\underline{s}(t-T). \tag{4.1}$$

Equation (4.1) is the same for all the models presented here and it will not be repeated again. Then in the second interval, [t-T+1:t-T] to [t:t-1], $t-T+1 \leq r \leq t$, employment $\underline{e}(r)$ is calculated from

$$\underline{e}(r) = \underline{b} + \Delta^*\underline{s}(r) + \sum_{w=t-T}^{r-1} \underline{s}^m(r,w) + \sum_{w=t-T}^{r-1} \underline{s}^s(r,w). \tag{4.2}$$

The last new changes are generated in [t:t-1] and thus in the interval [t+1:t] to [t+T+1:t+T], $t+1 \leq r \leq t+T+1$, $\Delta^*\underline{s}(r) = \underline{0}$. For this interval, the system is only redistributing itself, and these redistribution processes are losing momentum in themselves. Then for this interval

$$\underline{e}(r) = \underline{b} + \sum_{w=r-T-1}^{t} \underline{s}^{m}(r,w) + \sum_{w=r-T-1}^{t} \underline{s}^{s}(r,w) + \sum_{w=t-T}^{r-T-2} \underline{s}^{s}(w+T+1,w). \qquad (4.3)$$

In fact, at time r=t+1, the last term in equation (4.3) is out of range thus indicating that the system is at the point of balance between its initial growth and redistribution to time t, and its subsequent movement to stability from t+1 to t+T+1. Strictly speaking, an additional equation should be developed specifically for t+1, although this is not sufficiently different from equation (4.3) to warrant separate treatment. Figure 4.1 presents these processes diagrammatically and also serves to show the organisation of the simulation into three time intervals. Note that the central time period [t+1:t] indicates the change from growth to redistribution. At time t+T+2, equation (4.3) demonstrates that the system is composed entirely of stayers, and thus is in equilibrium.

Models for each of these processes were postulated in the last chapter where it was suggested that employment was related to population, and population to employment through a series of non-negative matrices reflecting the scale relations between these variables and their spatial interaction. For new change, service employment $\Delta^*\underline{s}(r)$ can be derived from the successive application of matrices $\underline{A}(\tau)$ and $\underline{B}(\tau)$ to the initial input $\underline{b}$: in effect, the process is one of calculating population from employment through the matrix $\underline{A}(\tau)$ and further employment from population through the matrix $\underline{B}(\tau)$, and it is in this sense that the population and employment equations are linked. Then

$$\Delta^*\underline{s}(r) = \underline{b} \prod_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau), \qquad (4.4)$$

and as it is assumed that $\Delta^*\underline{s}(r)<\Delta^*\underline{s}(r-1)$, it is clear that the matrix product $\underline{A}(\tau)\underline{B}(\tau)$ must be convergent in the Leontief sense (Gale, 1960).

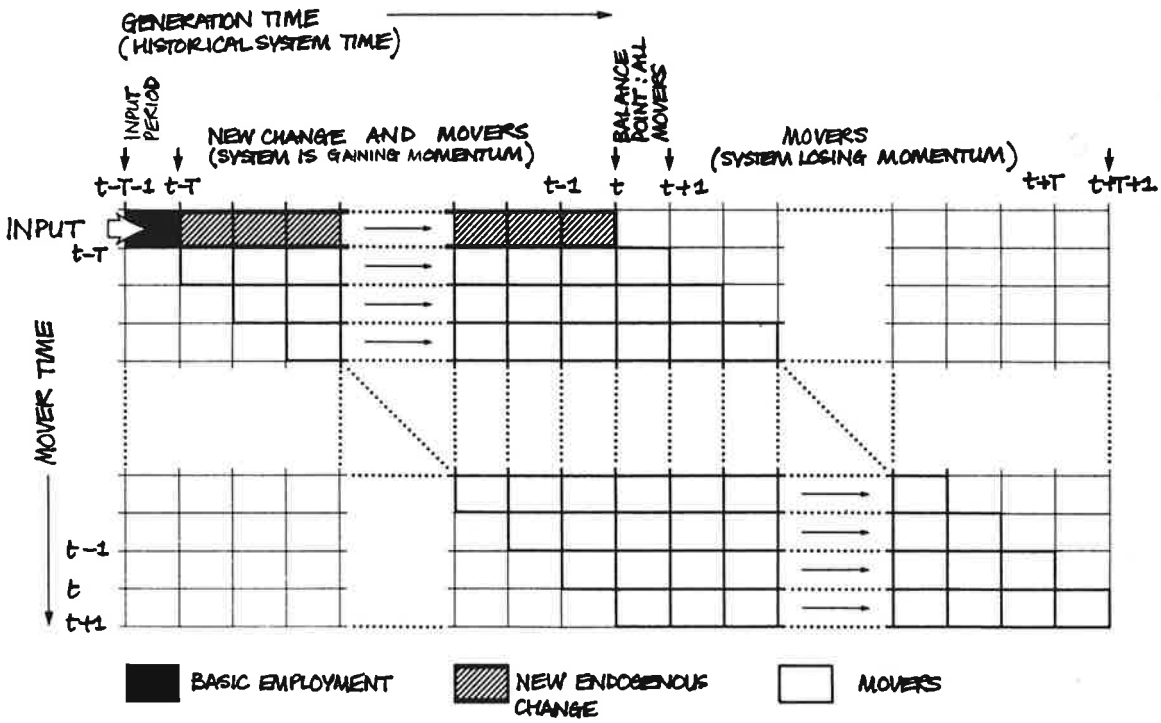



Figure 4.1: Temporal Structure of the Pseudo-Dynamic Model.

Movers are also modelled using a similar process which is lagged behind the process of new change but has a similar convergent character. The initial input $\underline{b}$ is redistributed using new scale and distribution incorporated in the matrices $\underline{\tilde{A}}(\tau)$ and $\underline{\tilde{B}}(\tau)$, and a proportion of activity $\underline{\alpha}(r,r-u)$ is subject to redistribution. $\underline{\alpha}(r,r-u)$ is an N x N scalar diagonal matrix where it is assumed that $\alpha_{ii}$ is constant. The index u is defined here as u=w-(t-T)=w-t+T, the amount of time between the generation of new activity and its origin. Movers are calculated from

$$\underline{s}^m(r,w) = \underline{b}\ \underline{\alpha}(r,r-u) \prod_{\tau=r-u}^{r} \underline{\tilde{A}}(\tau)\underline{\tilde{B}}(\tau) \ , \tag{4.5}$$

and stayers are then computed from the recurrence relation

$$\begin{aligned} \underline{s}^s(r,w) &= [\underline{s}^m(r-1,w)+\underline{s}^s(r-1,w)][\underline{I}-\underline{\alpha}(r,r-u)] \\ &= \underline{b} \sum_{\tau=w+1}^{r-1} \underline{\alpha}(\tau,\tau-u) \prod_{z=\tau-u}^{\tau} \underline{\tilde{A}}(z)\underline{\tilde{B}}(z) \prod_{v=\tau+1}^{r} [\underline{I}-\underline{\alpha}(v,v-u)] + \\ &\quad \underline{b} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z) \prod_{v=w+1}^{r} [\underline{I}-\underline{\alpha}(v,v-u)]. \end{aligned} \tag{4.6}$$

The amount of activity moving is controlled by the ratio matrix $\underline{\alpha}(r,r-u)$ which is dependent upon the system time r, and the generation time r-u.

In the original statement of this model, the recurrence relation used to calculate stayers was not expanded to the detail presented in equation (4.6). Here it is necessary to be more specific for in the typology of models to be developed below, it is necessary to examine the final configuration of the system at time t+T+1; hence the need for equation (4.6). Furthermore, the last term in equation (4.3) which concerns the activity which is stable, that is, stayers from previous times which are unchanging, needs elaboration. Equation (4.6) is not completely general for the stayers who actually remain stable as the model moves to

equilibrium, include the movers from the last significant time period who no longer move. Then

$$\underline{s}^{s}(w+T+1,w) = \underline{b} \sum_{\tau=w+1}^{w+T+1} \underline{\alpha}(\tau,\tau-u) \prod_{z=\tau-u}^{\tau} \tilde{\underline{A}}(z)\tilde{\underline{B}}(z) \prod_{v=\tau+1}^{w+T+1} [\underline{I}-\underline{\alpha}(v,v-u)]$$

$$+ \underline{b} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z) \prod_{v=w+1}^{w+T+1} [\underline{I}-\underline{\alpha}(v,v-u)]. \qquad (4.7)$$

As equation (4.7) does not depend upon r, it is constant for the appropriate point in time associated with equation (4.3).

Within the structure of this model, there are several elements which might vary and thus generate specific model types. In particular, the scale and distribution matrices $\underline{A}(\tau)$, $\underline{B}(\tau)$, $\tilde{\underline{A}}(\tau)$ and $\tilde{\underline{B}}(\tau)$ might take on different forms, as might the matrix $\underline{\alpha}(r,z)$ which controls the amount of activity which moves. In developing a typology of models based on specific forms for these elements, it is necessary to be fairly restrictive, that is, to assume that these elements are constant in different ways. By generating extreme cases of this type, it is possible to examine the limits of such a pseudo-dynamic model, but at the same time, some rather interesting structures emerge. In fact, these model types are fairly realistic in certain respects, and are useful in emphasising and simulating certain special processes. Many existing models which have a potential pseudo-dynamic interpretation can be generated in this manner, thus illustrating the richness of the typology. The main assumption adopted here which structures the following classification concerns the constancy of $\underline{\alpha}(r,z)$. It is assumed that $\underline{\alpha}(r,z) = \underline{\alpha}$ which is independent of mover and generation time, and that three model types are significant: $\underline{\alpha}=\underline{0}$ models, $\underline{\alpha}=\underline{I}$ models and $\underline{0}<\underline{\alpha}<\underline{I}$ models, thus implying models with no movers, models

with all activity distributed so far subject to redistribution, and models with partial but constant redistribution of the generated activity.

Within these three types, it is possible to make assumptions about the distribution of new activity and the redistribution of existing activity. Six cases are defined: first, the original model in which the distribution and redistribution effects vary through time; second, a model in which the redistribution effects are constant, that is $\underline{\tilde{A}}(\tau) = \underline{\tilde{A}}$ and $\underline{\tilde{B}}(\tau) = \underline{\tilde{B}}$; third, a model in which the distribution effects are constant, that is $\underline{A}(\tau) = \underline{A}$ and $\underline{B}(\tau) = \underline{B}$; fourth, a model in which both distribution and redistribution effects are constant; fifth a model in which distribution and redistribution effects are constant and equivalent; and sixth, a model in which the distribution effects are variable but equivalent, that is $\underline{A}(\tau) = \underline{\tilde{A}}(\tau)$ and $\underline{B}(\tau) = \underline{B}(\tau)$. In the sequel, $\underline{\alpha}=0$ models will first be developed followed by $\underline{\alpha}=\underline{I}$ models. A separate section is devoted to $\underline{\alpha}=\underline{I}$ models with equivalent but variable distribution and redistribution effects, and finally $\underline{0}<\underline{\alpha}<\underline{I}$ models are developed. In the next four sections, the model in equations (4.2) to (4.7) is first simplified according to the various assumptions concerning $\underline{\alpha}$ and then each of the six cases is developed where appropriate.

$\underline{\alpha}=\underline{0}$ MODELS: NO REDISTRIBUTION OF EXISTING ACTIVITY.

These models are by far the simplest in the typology for the lack of any redistribution-mover effects due to $\underline{\alpha}=\underline{0}$ collapses the model structure quite substantially. If there are no movers, it is intuitively obvious that the stayers will correspond to the new change generated so far which becomes immediately stable. Moreover, the only significant time interval

is from [t-T:t-T-1] to [t:t-1] for after that time, the system is stable. Substituting equations (4.4), (4.5) and (4.6), $\underline{e}(r)$, $t-T+1 \leq r \leq t$ into equation (4.2) gives

$$\underline{e}(r) = \underline{b} + \underline{b} \prod_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau) + \sum_{w=t-T}^{r-1} \underline{s}^{s}(r,w). \tag{4.8}$$

From the recurrence relation in equation (4.6), $\underline{s}^{s}(r,w)$ simplifies to

$$\underline{s}^{s}(r,w) = \underline{s}^{s}(w,w) = \underline{b} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z), \tag{4.9}$$

and thus equation (4.8) can be written as

$$\underline{e}(r) = \underline{b}\,[\underline{I} + \sum_{w=t-T}^{r} \prod_{\tau=t-T}^{w} \underline{A}(\tau)\underline{B}(\tau)]. \tag{4.10}$$

This is the same as Berechman's (1976) model in which he formulates the distribution matrices $\underline{A}(\tau)\underline{B}(\tau)$ using a non-stationary Markov process.

Only two of the six cases apply for this type of model and these relate to variable and constant distribution matrices. In the case of variable distribution matrices, there is little more to be said for the model is as given in equation (4.10). No further simplification is possible unless specific forms are adopted for $\underline{A}(\tau)\underline{B}(\tau)$ using some additional model structure, such as information-minimising (Batty and March, 1978) or a Markov process (Berechman, 1976; Stone, 1970). However, considerable simplification is possible if $\underline{A}(\tau)\ \underline{B}(\tau) = \underline{A}\ \underline{B}$, and then equation (4.10) becomes

$$\underline{e}(r) = \underline{b}\,[\underline{I} + \sum_{w=t-T}^{r} (\underline{A}\ \underline{B})^{w-t+T+1}]. \tag{4.11}$$

Under certain conditions, the matrix series in equation (4.11) is geometrically convergent in the sense that $\lim_{\tau\to\infty}(\underline{A}\ \underline{B})^{\tau} = \underline{0}$. For example, the assumptions about the form of $\underline{A}$ and $\underline{B}$ made in the last chapter in equations (3.19) and (3.20) would ensure convergence, and because it has been assumed that the process is, to all intents and purposes, convergent after T+1 increments of new change have been generated, then the series can be approximated by the inverse $(\underline{I} - \underline{A}\ \underline{B})^{-1}$ when r=t. That is

$$\underline{e}(t) = \underline{b}[\underline{I} - \underline{A}\ \underline{B}]^{-1} . \qquad (4.12)$$

This is, of course, a well-known form. It is one version of the so-called Lowry model derived simultaneously and independently by Harris (1966) and Garin (1966) from Lowry's (1964) original Pittsburgh model. It represents the simplest spatial equivalent of the input-output model presented in Chapter 2. Equation (4.12) also represents the model from which much of this analysis has been derived and it is the basis of the dynamic framework developed in Chapter 3, and already used for a model of the Reading subregion by the author (Batty, 1976). As such, it is the simplest of all the models presented here and has perhaps the least interest due to its well-known form and widespread application.

## $\underline{\alpha}=\underline{I}$ MODELS: COMPLETE REDISTRIBUTION OF EXISTING ACTIVITY.

In this class of model, all the activity generated so far is reallocated in the subsequent period of time. In essence, all existing activity is moved, no activity remains stable. The simplification of the original model is not as substantial as in the previous set of models, although it is still considerable. For $t-T+1 \leq r \leq t$, the employment equation is based solely on the input, on new change and on movers

$$\underline{e}(r) = \underline{b} + \underline{b} \prod_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau) + \underline{b} \sum_{v=t-T+1}^{r} \prod_{\tau=v}^{r} \tilde{\underline{A}}(\tau)\, \tilde{\underline{B}}(\tau). \qquad (4.13)$$

Note that the term which models movers in equation (4.13) demonstrates that all the activity generated up until r-1 is reallocated. From $t+1 \leq r \leq t+T+1$, the model equation becomes

$$\underline{e}(r) = \underline{b}+\underline{b} \sum_{v=r-T}^{t+1} \prod_{\tau=v}^{r} \tilde{\underline{A}}(\tau)\tilde{\underline{B}}(\tau) + \underline{b} \sum_{w=t-T}^{r-T-2} \prod_{z=t+1}^{w+T+1} \tilde{\underline{A}}(z)\tilde{\underline{B}}(z), \qquad (4.14)$$

and it is obvious from the range of summation and multiplication in the above equations that the movers are lagged one period behind the original generation of new change.

A clearer demonstration of the movement of all the existing stock according to new distribution and redistribution matrices at time r can be made in relation to equation (4.13). Equation (4.13) can be rewritten as

$$\underline{e}(r) = \underline{b}\, \{\underline{I} + \prod_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau)+[\underline{I} + \sum_{v=t-T+1}^{r-1} \prod_{\tau=v}^{r-1} \tilde{\underline{A}}(\tau)\tilde{\underline{B}}(\tau)]\tilde{\underline{A}}(r)\tilde{\underline{B}}(r)\}. \qquad (4.15)$$

From equation (4.15), it is clear that the original pattern is lost immediately and that at each time period, the original sequence of generation begins to be reallocated afresh. By time t+1, all the activity is generated as new change has been reallocated. In short, the original distributions have been destroyed, and it might be argued that there is no purpose to including them. But in models of this type as will be evident later, there is often positive feedback at each stage from the state of the system to the form of the distribution and redistribution matrices used at the succeeding stage, and thus the processes as specified above are all essential.

No further analysis of the case where the distribution and redistribution

effects vary through time is possible other than that already presented in equations (4.13) to (4.15) but substantial simplification occurs for the second case in which the redistribution matrices are constant over the simulation. Taking equation (4.15) and setting $\underline{\tilde{A}}(\tau) = \underline{\tilde{A}}$ and $\underline{\tilde{B}}(\tau) = \tilde{B}$ yields the following form

$$\underline{e}(r) = \underline{b}\{\underline{I}+ \prod_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau)+[\underline{I}+ \sum_{v=t-T+1}^{r-1} (\underline{\tilde{A}}\ \underline{\tilde{B}})^{r-v}]\underline{\tilde{A}}\ \underline{\tilde{B}}\}$$

$$= \underline{b}\{\underline{I}+ \prod_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau)+[\underline{I}-(\underline{\tilde{A}}\ \underline{\tilde{B}})^{r-t+T}][\underline{I}-\underline{\tilde{A}}\ \underline{\tilde{B}}]^{-1}\underline{\tilde{A}}\ \underline{\tilde{B}}\}. \qquad (4.16)$$

Using the series simplification for movers redistributed by the constant matrices $\underline{\tilde{A}}\ \underline{\tilde{B}}$ leads to stability at r=t+1. Then from equation (4.14)

$$\underline{e}(t+1) = \underline{b}\{\underline{I}+[\underline{I}-(\underline{\tilde{A}}\ \underline{\tilde{B}})^{T+1}][\underline{I}-\underline{\tilde{A}}\ \underline{\tilde{B}}]^{-1}\underline{\tilde{A}}\ \underline{\tilde{B}}\} ,$$

$$= \underline{b}\{\underline{I}+\underline{\tilde{A}}\ \underline{\tilde{B}}[\underline{I}-\underline{\tilde{A}}\ \underline{\tilde{B}}]^{-1}\}$$

$$= \underline{b}[\underline{I}-\underline{\tilde{A}}\ \underline{\tilde{B}}]^{-1} . \qquad (4.17)$$

Equation (4.17) is identical to equation (4.12) with $\underline{\tilde{A}}\ \underline{\tilde{B}}$ replacing $\underline{A}\ \underline{B}$ and it leads to the obvious point that if $\underline{\tilde{A}}\ \underline{\tilde{B}}$ is known before the start of the simulation, there is little point to the model process as specified in equation (4.16). However, there is a use for the model if $\underline{\tilde{A}}\ \underline{\tilde{B}}$ is determined after $\underline{e}(t-T)$ has been calculated or if $\underline{\alpha}=\underline{0}$ until r=R when $\underline{\alpha}=\underline{I}$ and $\underline{\tilde{A}}\ \underline{\tilde{B}}$ is formed. In this sense, there is no guarantee that the complete mover process need be initiated at the start of the simulation, and it would only be operative if some particular condition were met at $r \pm R$. In this situation, equation (4.16) would still be applicable but only for $r \geq R$. For $r \geq R$ equation (4.10) from the previous class of models would hold.

The case in which $\underline{A}(\tau) = \underline{A}$ and $\underline{B}(\tau) = \underline{B}$ does not lead to any significant

form, and in the case where this constancy in distribution is combined with the above constancy in redistribution, the model form is similar to that in equations (4.16) and (4.17) with the matrix product $\Pi_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau)$ being replaced by $(\underline{A}\ \underline{B})^{r-t+T+1}$. The fifth case where the distribution and redistribution is constant and equivalent is trivial for redistribution does not change the original distribution and thus the model is identical to the Garin-Harris version of Lowry's model in equations (4.11) and (4.12).

The sixth and final case in which $\underline{A}(\tau) = \tilde{\underline{A}}(\tau)$ and $\underline{B}(\tau) = \tilde{\underline{B}}(\tau)$ is by far the most interesting for it involves the same process for redistribution as distribution but with the additional point that the lag between distribution and redistribution is one time period. Thus new activity is generated and distributed in the same way that existing activity is regenerated and redistributed. A version of this model has been developed recently by Baxter and Williams (1975) in a somewhat different guise, and this model type is so important that a separate section for its elaboration is warranted. These types of models are henceforth referred to collectively as Baxter-Williams type models.

BAXTER-WILLIAMS TYPE MODELS.

Assuming the equivalence of the distribution and redistribution matrices defined above, the employment equation for $t-T+1 \leq r \leq t$ can be derived from equation (4.15). Then

$$\underline{e}(r) = \underline{b}\{\underline{I}+ \prod_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau)+[\underline{I}+ \sum_{v=t-T+1}^{r-1} \prod_{\tau=v}^{r-1} \underline{A}(\tau)\ \underline{B}(\tau)]\underline{A}(r)\underline{B}(r)\},$$

$$= \underline{b}\{\underline{I}+ \sum_{v=t-T}^{r} \prod_{\tau=v}^{r} \underline{A}(\tau)\ \underline{B}(\tau)\}. \qquad (4.18)$$

Equation (4.18) is very similar to Berechman's (1976) version of the Lowry model but there is a significant difference in the way the matrix product is taken. In essence, equation (4.18) is a backwards version of Berechman's model in which basic employment and its repercussions in terms of new change are being reallocated using the distribution matrices taken from r back to t-T.

The essential structure of this model can be made apparent by rewriting equation (4.18) as

$$\underline{e}(r) = \underline{b}\{\underline{I}+[\underline{I}+\sum_{v=t-T}^{r-1}\prod_{\tau=v}^{r-1}\underline{A}(\tau)\underline{B}(\tau)]\underline{A}(r)\underline{B}(r)\},$$

$$= \underline{b} + \underline{e}(r-1)\underline{A}(r)\underline{B}(r). \qquad (4.19)$$

Equation (4.19) shows that all the employment generated so far is reallocated in the subsequent time period together with the relevant new change. This equation is the one derived using a different argument by Baxter and Williams (1975) who develop it for purposes of easing the calibration problem of the urban model proposed by Echenique, Crowther and Lindsay (1969). Moreover, as will be demonstrated in later chapters the model is eminently suited to dealing with locational constraints in a manner not applied so far. The original model derived by Baxter and Williams (1975) is presented in Appendix 1 from which it is clear that their model has a pseudo-dynamic form.

It is now necessary to examine the form of the model after time t when new activity is no longer being generated. Equation (4.14) holds for this process and it is of interest to examine the specific structure of activity in the last time period of the simulation, that is, when r=t+T+1. Then

$$\underline{e}(t+T+1) = \underline{b} + \underline{b} \prod_{\tau=t+1}^{t+T+1} \underline{A}(\tau)\underline{B}(\tau) + \underline{b} \sum_{w=t-T}^{t-1} \prod_{z=t+1}^{w+T+1} \underline{A}(z)\underline{B}(z),$$

$$= \underline{b}\{\underline{I} + \sum_{w=t-T}^{t} \prod_{z=t+1}^{w+T+1} \underline{A}(z)\underline{B}(z)\}. \tag{4.20}$$

Equation (4.20) is extremely interesting for it indicates that the final distribution of activity is distributed according to the $\underline{\alpha}=\underline{0}$ type model but with the sequence of distribution taken from the time when all the activity in the system has been first generated to the time when the last activity generated has been regenerated and redistributed. In short, this model involves a backwards progression of redistribution in the interval $t-T\leq r\leq t$ and a forwards progression in $t+1\leq r\leq t+T+1$, with the added point that the forwards progression is thoroughly dependent for its form on the backwards progression.

Because equation (4.20) has the structure of an $\underline{\alpha}=\underline{0}$ model, the results pertaining to the earlier section on these models apply. The only significant simplification for these models results when $\underline{A}(z)\underline{B}(z)$ is constant. If it is assumed that $\underline{A}(z)\underline{B}(z)$ becomes constant after the last increment of new change is generated, that is $\underline{A}(\tau)\underline{B}(\tau) = \underline{A}(t)\underline{B}(t)$, $\tau\geq t$, then equation (4.20) simplifies to

$$\underline{e}(t+T+1) = \underline{b}\ \{\underline{I}+ \sum_{w=t-T}^{t} [\underline{A}(t)\underline{B}(t)]^{w-t+T+1}\}, \tag{4.21}$$

which in turn can be approximated by

$$\underline{e}(t+T+1) = \underline{b}[\underline{I}-\underline{A}(t)\underline{B}(t)]^{-1}. \tag{4.22}$$

Note that equation (4.22) can also be derived from slightly different considerations as demonstrated in Appendix 1. This completes the discussion of $\underline{\alpha}=\underline{I}$ models although these will be introduced again in later chapters in relation to the development of the procedures to handle

locational constraints outlined below.

## $\underline{\alpha}$ CONSTANT MODELS: PARTIAL REDISTRIBUTION OF EXISTING ACTIVITY.

The class of models generated when $\underline{\alpha}$ is assumed to be constant over time and space, that is, $\underline{0}<\underline{\alpha}<\underline{I}$, is the most complicated of the three special cases generated by adopting a constant form for $\underline{\alpha}$. In Chapter 3, this type of model was used to demonstrate the various rates at which the existing stock turned over and there it was shown that some simplification of the original pseudo-dynamic model form is possible. It is worthwhile developing the simplification explicitly for the time intervals $t-T+1\leq r\leq t$ and $t+1\leq r\leq t+T+1$, and from these equations, further simplification is then possible in terms of the six cases relating to distribution and redistribution. For $t-T+1\leq r\leq t+1$, $\underline{e}(r)$ is derived by substituting equations (4.4), (4.5) and (4.6) into (4.2) and simplifying. Then

$$\begin{aligned}\underline{e}(r) = {} & \underline{b} \\ & + \underline{b} \prod_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau) \\ & + \underline{b}\,\underline{\alpha} \sum_{v=t-T+1}^{r} \prod_{\tau=v}^{r} \tilde{\underline{A}}(\tau)\tilde{\underline{B}}(\tau) \\ & + \underline{b} \sum_{w=t-T}^{r-1} \{[\underline{\alpha} \sum_{\tau=w+1}^{r-1} (\underline{I}-\underline{\alpha})^{r-\tau} \prod_{z=\tau-u}^{\tau} \tilde{\underline{A}}(z)\tilde{\underline{B}}(z)] \\ & + [(\underline{I}-\underline{\alpha})^{r-w} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)] \}. \end{aligned} \tag{4.23}$$

Note that equation (4.23) is organised to show the input on the first line, the new change on the second, movers on the third and stayers on the fourth and subsequent lines.

This form is also adopted for $\underline{e}(r)$ in the period $t+1 \leq r \leq t+T+1$

$$
\begin{aligned}
\underline{e}(r) = \underline{b} & \\
& + \underline{b}\,\underline{\alpha} \sum_{v=r-T}^{t+1} \prod_{\tau=v}^{r} \tilde{\underline{A}}(\tau)\tilde{\underline{B}}(\tau) \\
& + \underline{b} \sum_{w=r-T-1}^{t} \{[\underline{\alpha} \sum_{\tau=w+1}^{r-1} (\underline{I}-\underline{\alpha})^{r-\tau} \prod_{z=\tau-u}^{\tau} \tilde{\underline{A}}(z)\tilde{\underline{B}}(z)] + \\
& \quad [(\underline{I}-\underline{\alpha})^{r-w} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)\ ]\} \\
& + \underline{b} \sum_{w=t-T}^{r-T-2} \{[\underline{\alpha} \sum_{\tau=w+1}^{w+T+1} (\underline{I}-\underline{\alpha})^{w+T+1-\tau} \prod_{z=\tau-u}^{\tau} \tilde{\underline{A}}(z)\tilde{\underline{B}}(z)] + \\
& \quad [(\underline{I}-\underline{\alpha})^{T+1} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)]\}, \qquad (4.24)
\end{aligned}
$$

where the first line is the input, the second the activity which is still moving, the third and fourth the stayers associated with those still moving and fifth and sixth the stayers who form part of the stable equilibrium. From equations (4.23) and (4.24), further simplifications can now be developed for the six cases relating to distribution and redistribution.

For the case where the distribution and redistribution matrices are distinct and time dependent, no simplification is possible although equations (4.23) and (4.24) provide forms through which the proportion of existing activity remaining can be computed. It is in the second case in which the redistribution matrices are assumed constant that most simplification results, that is, when $\tilde{\underline{A}}(\tau) = \tilde{\underline{A}}$ and $\tilde{\underline{B}}(\tau) = \tilde{\underline{B}}$. In this case, it is worthwhile looking at the appropriate forms of equations (4.23) and (4.24), and also at the situation where $r=t+T+1$. Then for $t-T+1 \leq r \leq t$,

$$
\begin{aligned}
\underline{e}(r) = \underline{b} & \\
& + \underline{b} \prod_{\tau=t-T}^{r} \underline{A}(\tau)\underline{B}(\tau) \\
& + \underline{b}\ \underline{\alpha}\ [\underline{I}-(\tilde{\underline{A}}\ \tilde{\underline{B}})^{r-t+T}]\tilde{\underline{A}}\ \tilde{\underline{B}}\ [\underline{I}-\tilde{\underline{A}}\ \tilde{\underline{B}}]^{-1} \\
& + \underline{b} \sum_{w=t-T}^{r-1} \{[\underline{I}-(\underline{I}-\underline{\alpha})^{r-w-1}](\tilde{\underline{A}}\ \tilde{\underline{B}})^{w-t+T+1} \quad + \\
& \quad [(\underline{I}-\underline{\alpha})^{r-w-1} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)]\}(\underline{I}-\underline{\alpha}) \ . \qquad (4.25)
\end{aligned}
$$

Note the simplification in the movers in which the series can be summarised by the appropriate part of a geometrically convergent series, and the simplification in the mover components of the stayers where the regeneration (mover) ratio can be simplified in a similar manner.

For the time interval $t+1 \leq r \leq t+T+1$, greater simplification is possible

$$
\begin{aligned}
\underline{e}(r) = \underline{b} & \\
& + \underline{b}\ \underline{\alpha}(\tilde{\underline{A}}\ \tilde{\underline{B}})^{r-t}[\underline{I}-\tilde{\underline{A}}\ \tilde{\underline{B}}]^{-1} \\
& + \underline{b} \sum_{w=r-T-1}^{t} \{[\underline{I}-(\underline{I}-\underline{\alpha})^{r-w-1}]\ (\tilde{\underline{A}}\ \tilde{\underline{B}})^{w-t+T+1} \quad + \\
& \quad [(\underline{I}-\underline{\alpha})^{T+1} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)]\} \\
& + \underline{b}\ \{[\underline{I}-(\underline{I}-\underline{\alpha})^{T+1}][\underline{I}-(\tilde{\underline{A}}\ \tilde{\underline{B}})^{r-t-1}]\tilde{\underline{A}}\ \tilde{\underline{B}}\ [\underline{I}-\tilde{\underline{A}}\ \tilde{\underline{B}}]^{-1} \quad + \\
& \quad (\underline{I}-\underline{\alpha})^{T+1} \sum_{w=t-T}^{r-T-2} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)]\} \ . \qquad (4.26)
\end{aligned}
$$

Perhaps the most interesting equation characterising this model type relates to the situation at $r=t+T+1$ or $r=t+T+2$ which is the stable situation. Then from equation (4.26) with $r=t+T+2$

$$\underline{e}(t+T+2) = \underline{b} + \underline{b}[\underline{I}-(\underline{I}-\underline{\alpha})^{T+1}]\,\tilde{\underline{A}}\,\tilde{\underline{B}}\,[\underline{I}-\tilde{\underline{A}}\,\tilde{\underline{B}}]^{-1}$$

$$+ \underline{b}[(\underline{I}-\underline{\alpha})^{T+1} \sum_{w=t-T}^{t} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)]. \qquad (4.27)$$

In fact, the real structure of equation (4.27) can be displayed by rearranging

$$\underline{e}(t+T+2) = \underline{b} + \underline{b}\,\tilde{\underline{A}}\,\tilde{\underline{B}}[\underline{I}-\tilde{\underline{A}}\,\tilde{\underline{B}}]^{-1}$$

$$- (\underline{I}-\underline{\alpha})^{T+1}\{\tilde{\underline{A}}\,\tilde{\underline{B}}\,[\underline{I}-\tilde{\underline{A}}\,\tilde{\underline{B}}]^{-1} - \sum_{w=t-T}^{t} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)\}. \qquad (4.28)$$

Clearly the equilibrium situation can be interpreted as a structure of activity which is composed of a proportion of the existing pattern and the rest the new pattern due to constant redistribution. Equation (4.28) is suitable if $(\underline{I}-\underline{\alpha})^{T+1}$ is near to $\underline{0}$ because it indicates that the structure can be seen as largely due to the redistribution less a proportion $(\underline{I}-\underline{\alpha})^{T+1}$ of the difference between the new and original distributions.

Equations (4.23) to (4.28) present a model which is probably of limited operational interest, for the structure does not give much opportunity for positive feedback from distribution to redistribution and vice versa. Nevertheless, it might be of some importance where information is known about variable distribution and constant redistribution processes. Furthermore, these equations are useful for generating other types of $\underline{0}<\underline{\alpha}<\underline{I}$ model, and also as another means of deriving the appropriate $\underline{\alpha}=\underline{0}$ and $\underline{\alpha}=\underline{I}$ types. For the case where the distribution matrices are constant and the redistribution variable, equations (4.23) and (4.24) hold with the product term characterising the original distribution replaced by $(\underline{A}\,\underline{B})^{r-t+T+1}$ and similar power functions in the rest of the two equations. This model may be of some use because feedback from distribution to redistribution

is possible and thus the model can potentially handle constraints on its dynamic process.

Where both the distribution and redistribution matrices are constant, the model is similar to that outlined in equations (4.25) to (4.28) with power terms replacing the distribution matrix products. In fact, although the model is of marginal interest for it would only pertain to the case where the distribution and redistribution, generation and regeneration were constant and independent of time, the final configuration of activities is interesting. From equation (4.28) with $\underline{A}(\tau) = \underline{A}$ and $\underline{B}(\tau) = \underline{B}$, the structure at t+T+2 is

$$\underline{e}(t+T+2) = \underline{b} + \underline{b}\,\tilde{\underline{A}}\,\tilde{\underline{B}}\,[\underline{I}-\tilde{\underline{A}}\,\tilde{\underline{B}}]^{-1} - (\underline{I}-\underline{\alpha})^{T+1}\{\tilde{\underline{A}}\,\tilde{\underline{B}}[\underline{I}-\tilde{\underline{A}}\,\tilde{\underline{B}}]^{-1} - \underline{A}\,\underline{B}[\underline{I}-\underline{A}\,\underline{B}]^{-1}\}. \qquad (4.29)$$

Equation (4.29) indicates the intuitively obvious result that the final configuration is completely independent of historical time, apart from the length of the life of the distribution-redistribution process T. The fifth case in which distribution is equivalent to redistribution and constant, that is, $\underline{A}=\tilde{\underline{A}}$ and $\underline{B}=\tilde{\underline{B}}$, generates a model which is the same as the $\underline{\alpha}=\underline{0}$ type model. Using this assumption in equation (4.29), it is clear that the final configuration is the same as that produced by a Lowry model [see equations (4.12), (4.17) and (4.22)].

One final case remains to be dealt with and that involves the $\underline{0}<\underline{\alpha}<\underline{I}$ equivalent of the Baxter-Williams type models. These models are worth detailing because they are particularly relevant to the treatment of constraints on the locational process developed in the next section. Assuming $\underline{A}(\tau)=\tilde{\underline{A}}(\tau)$ and $\underline{B}(\tau)=\tilde{\underline{B}}(\tau)$, the model can be written out for each time interval as in previous cases. Then for $t-T+1\leq r\leq t$

$$\underline{e}(r) = \underline{b}+\underline{b}\{\underline{\alpha}+\underline{\alpha} \sum_{v=t-T+1}^{r-1} \prod_{\tau=v}^{r-1} \underline{A}(\tau)\underline{B}(\tau) + \prod_{\tau=t-T}^{r-1} \underline{A}(\tau)\underline{B}(\tau)\}\underline{A}(r)\underline{B}(r)$$

$$+ \underline{b} \sum_{w=t-T}^{r-1} \{[\underline{\alpha} \sum_{\tau=w+1}^{r-1} (\underline{I}-\underline{\alpha})^{r-\tau} \prod_{z=\tau-u}^{\tau} \underline{A}(z)\underline{B}(z)] +$$

$$[(\underline{I}-\underline{\alpha})^{r-w} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)]\}. \qquad (4.30)$$

The main characteristic of equation (4.30) relates to the fact that the model is based on the usual backwards process for the redistribution of movers from the existing activity and a more complicated expression of the usual form modelling the stayers. For $t+1 \leq r \leq t+T+1$, the equation is

$$\underline{e}(r) = \underline{b} + \underline{b}\{\underline{\alpha}+\underline{\alpha} \sum_{v=r-T+1}^{t+1} \prod_{\tau=v}^{r-1} \underline{A}(\tau)\underline{B}(\tau)\}\underline{A}(r)\underline{B}(r)$$

$$+ \underline{b} \sum_{w=r-T-1}^{t} \{[\underline{\alpha} \sum_{\tau=w+1}^{r-1} (\underline{I}-\underline{\alpha})^{r-\tau} \prod_{z=\tau-u}^{\tau} \underline{A}(z)\underline{B}(z)] +$$

$$[(\underline{I}-\underline{\alpha})^{r-w} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)]\}$$

$$+ \underline{b} \sum_{w=t-T}^{r-T-2} \{[\underline{\alpha} \sum_{\tau=w+1}^{w+T+1} (\underline{I}-\underline{\alpha})^{w+T+1-\tau} \prod_{z=\tau-u}^{\tau} \underline{A}(z)\underline{B}(z)] +$$

$$[(\underline{I}-\underline{\alpha})^{T+1} \sum_{z=t-T}^{w} \underline{A}(z)\underline{B}(z)]\}. \qquad (4.31)$$

The process is only of additional interest if it is assumed that $\underline{A}(r)\underline{B}(r)$ become constant as in the $\underline{\alpha}=\underline{I}$ version of this model.

Then at the time $r=t$, assume that $\underline{A}(\tau)=A(t)$, $\underline{B}(\tau)=B(t), \tau \geq t$ , and the final configuration of activity $t+T+2$ is then given by

$$\underline{e}(t+T+2) = \underline{b} + \underline{b} \sum_{w=t-T}^{t} [\underline{\alpha} \sum_{\tau=w+1}^{t} (\underline{I}-\underline{\alpha})^{w+T+1-\tau} \prod_{z=\tau-u}^{\tau} \underline{A}(z)\underline{B}(z)] +$$

$$\underline{b}\ \underline{A}(t)\underline{B}(t)\{[\underline{I}-\underline{A}(t)\underline{B}(t)]^{-1}-(\underline{I}-\underline{\alpha})[\underline{I}-\underline{A}(t)\underline{B}(t)+\underline{\alpha}\ \underline{A}(t)\underline{B}(t)]^{-1}\} +$$

$$\underline{b}(\underline{I}-\underline{\alpha})^{T+1} \sum_{w=t-T}^{t} \prod_{z=t-T}^{w} \underline{A}(z)\underline{B}(z). \qquad (4.32)$$

Equation (4.32) is in three parts: the first line gives the input and the stayers associated with all the time dependent movement up to time t, the second line the stayers associated with the time independent movement after t, and the third line the stayers associated with the original distribution of new change. As the three major types of model generated from assumptions of constancy in the generation-regeneration, distribution-redistribution elements of the model have now been presented, it is worthwhile exploring the way in which the mover processes can be used to incorporate constraints and controls on the various states predicted by the model. Such a discussion is a necessary preliminary to an application of one of these types of pseudo-dynamic model which is to be presented in the next chapter.

## CONTROLS AND CONSTRAINTS ON THE PSEUDO-DYNAMIC PROCESS.

In the original dynamic framework from which the idea of a pseudo-dynamic model was derived in Chapter 3, the concept of redistribution by movement of existing activities was regarded as a central process characterising the behaviour of urban systems. The argument was based on the notion that as conditions in the system change, activity which already exists and has originally been generated from changes in the system's environment, must be redistributed to meet the changed conditions. There are many examples of this type of process: the invasion and succession of different land uses in the city, shifts in land use due to changes in the pattern of

accessibility due in turn to changes in transport technology and infrastructure, changes in land use due to obsolescence, redevelopment and so on.

Indeed, there are many studies which show that a large proportion of all the change in a city (greater than 75 percent, say) is due to redistribution, rather than new growth or decline which depends on changes in the system's environment. It is possible to interpret these processes of redistribution as mechanisms by which the system keeps itself on course by adapting to its own changed circumstances, and thus the idea of the mover processes 'controlling' or 'constraining' the form of the system seems attractive. At least in this context, the idea of control if not explicitly related to the powerful results available in control theory, is a suggestive means of demonstrating how these types of models might be constrained in various ways.

The processes which characterise the distribution and redistribution of activity in the pseudo-dynamic model, are based on the ratio $\underline{\alpha}$ which regenerates a proportion of the activity already generated, and the interaction matrices $\underline{A}(\tau)$, $\underline{B}(\tau)$, $\tilde{\underline{A}}(\tau)$ and $\tilde{\underline{B}}(\tau)$ which determine the allocation and reallocation of activity. The implication is that these elements depend upon the state of the system at any time r; that is, that these elements are determined in some way by the state of the system in relation to some set of prior conditions or targets which the system must meet. In short, there is feedback from the state of the system at time r to the means by which activity is distributed and/or redistributed at time r+1.

Taking the employment equation $\underline{e}(r)$ which is given in aggregate form as

equation (4.2), the following scheme illustrates the notion of feedback through the mover process in the quest to get the system to reach some target. Equation (4.2) is first repeated

$$\underline{e}(r) = \underline{b} + \Delta^*\underline{s}(r) + \sum_{w=t-T}^{r-1} \underline{s}^m(r,w) + \sum_{w=t-T}^{r-1} \underline{s}^s(r,w), \qquad [(4.2)]$$

and an analogous equation for population also exists. Then from the population $\underline{p}(r)$ and employment $\underline{e}(r)$ at time r, the system is evaluated in terms of the targets and constraints on $\underline{p}(r)$ and $\underline{e}(r)$ to be met, called $\underline{c}^p(r)$ and $\underline{c}^e(r)$ respectively. Furthermore the existing means of re-distributing activity through $\tilde{\underline{A}}(r)$ and $\tilde{\underline{B}}(r)$ are to be modified on the basis of the state of the system in relation to its constraints, as is the mover ratio matrix $\underline{\alpha}(r,w)$. Then some typical feedback control functions might be of the form

$$\tilde{\underline{A}}(r+1) = f^1[\underline{p}(r), \underline{c}^p(r), \underline{e}(r), \underline{c}^e(r), \tilde{\underline{A}}(r)] ,$$
$$\tilde{\underline{B}}(r+1) = f^2[\underline{p}(r), \underline{c}^p(r), \underline{e}(r), \underline{c}^e(r), \tilde{\underline{B}}(r)] , \quad \text{and}$$
$$\underline{\alpha}(r+1,w) = f^3[\underline{p}(r), \underline{c}^p(r), \underline{e}(r), \underline{c}^e(r), \underline{\alpha}(r,w)]$$

From equation (4.5) of the model, it is clear that movers $\underline{s}^m(r+1,w)$ during the period [r+1:r] are a function of $\tilde{\underline{A}}(r+1)\tilde{\underline{B}}(r+1)$ and $\underline{\alpha}(r+1,w)$ and thus it follows that employment $\underline{e}(r+1)$ is a function of the state of the system at time r as well as other elements specified exogenously.

This discussion shows that although the idea of feedback control is suggestive, it is mathematically quite tricky: in this example, it is certainly nonlinear and probably discontinuous as will be demonstrated in later chapters, and thus the absence of any simple linear feedback control is likely to make the mathematics cumbersome and somewhat inelegant. Moreover, the model will probably have to be solved using some form of

iteration. Indeed in the $\underline{\alpha}=\underline{0}$ versions of this model in which there are no movers, constraints can still be incorporated by modifying the distribution matrices $\underline{A}(\tau)$ and $\underline{B}(\tau)$ in a similar fashion to that sketched above, although in such a case, some iteration on these matrices is likely (Batty, 1976; Berechman, 1976).

The emphasis adopted in this and the previous chapter has been on interpreting the model as a mainly spatial device to successively allocate and reallocate activity through time. In fact, the scale component which controls the amount which can be generated or regenerated, has been regarded as exogenous and independent of time, although it is possible to regard this as time dependent. Therefore, the treatment of constraints will be solely related to the spatial dimensions of this model despite the fact that the amount of activity could be controlled or constrained in an endogenous fashion. Two distinct types of constraint on spatial allocation can be recognised: first, constraints on the amount of activity locating in any zone, and second, constraints on the pattern of distribution or interaction between activities over zones.

The first type of constraint dealing with location can be specified in the manner described above in which the targets are set by the prespecified constraint vectors $\underline{c}^e(r)$ and $\underline{c}^p(r)$. In several applications of the $\underline{\alpha}=\underline{0}$ models, constraints have been set up in this fashion and met by solving the process from t-T to t several times in the quest to find a solution which satisfied the constraints, or by iteration within each time period [r+1:r] to satisfy the constraints in a pseudo-dynamic sense. The first type of constraint method has been used in various models by the author (Batty, 1976), the second type by Echenique, Crowther and Lindsay

(1969) in their Reading model and in a more extensive way, by Feo, Herrera, Riquezes and Echenique (1975) in their Caracas model.

In the next chapter the use of constraints will be restricted to locational constraints whereas in Chapters 6 to 9, the idea of constraints on the interaction process will be explored. Constraints on interaction which involve finding a solution to the model which generates a pattern of interaction consistent with some prior information about the form of the interaction pattern, relate to the process of estimating parameters of interaction. In essence, such constraints reflect the process of calibrating the model's distribution and redistribution matrices (which in themselves are based on submodels of interaction) to meet certain interaction criteria relating to trip frequency, average trip lengths and so on. In existing $\underline{\alpha}=\underline{0}$ (Lowry) models, such calibration has been static in that the interaction parameters have thus been estimated in a static sense for the whole process. As will be demonstrated in Chapter 6, new perspectives on the structure of these kinds of model are opened up by treating the calibration of a pseudo-dynamic model as a dynamic process in its own right.

The great advantage to using the mover processes as mechanisms for enabling the model to meet certain targets or constraints, is that it is always assured that such constraints will be met. In the pseudo-dynamic model, there are T+1 time periods associated with the generation of new change, and in each of these time periods after the first, a process of redistributing the whole sequence is possible. This leads to T+1 sequences of moves, in fact, the whole of the activity generated and distributed by the model can be regenerated and redistributed T+1 times. Therefore whenever a constraint is violated at time r, $t-T\leq r\leq t$, a mover sequence reallocating all the activity begins at the next time period [r+1:r]. Moreover, because the

reallocation continues for T+1 time periods, the constraints can still be checked at every period up to t+T+1.

It is intuitively obvious that if the process of constraint is well specified in terms of the interaction matrices, the constraints will always be met. However it is worth pointing out that the original pseudo-dynamic form involves considerably more computation than any of the simpler existing versions of this model such as the $\underline{\alpha}=\underline{0}$ models where constraint processes are dealt with in a fairly arbitrary manner. Moreover, there are certain special forms of pseudo-dynamic model in which constraints can be handled in a much more satisfactory way than at present, and some of these will be presented in the next chapter.

CONCLUSIONS.

In terms of the previous classification of models according to the form of the mover matrix $\underline{\alpha}(r,w)$, it is clear that two general types of mover process can be adopted in dealing with constraints: complete re-distribution, $\underline{\alpha}(r,w)=\underline{I}$, or partial redistribution, $\underline{0}<\underline{\alpha}(r,w)<\underline{I}$. Within this division, it is possible to consider mover processes which involve the whole sequence of activity as in the models presented so far, or processes where only part of the activity is redistributed; that is, where only certain increments of activity within the total process are moved. Furthermore, it seems logical that further distribution and redistribution would both be affected by the violation of locational constraints. Thus in the next chapter, it is assumed that constraints are handled using Baxter-Williams type models in which $\underline{A}(\tau)=\underline{\tilde{A}}(\tau)$ and $\underline{B}(\tau)=\underline{\tilde{B}}(\tau)$. Many other varieties of schema are possible for handling locational constraints and one of the great advantages of the pseudo-dynamic form

is its flexibility in this regard. In the next chapter these constraint procedures will be discussed first in terms of the complete redistribution of activity and second in terms of partial redistribution.

# CHAPTER 5.

# LOCATIONALLY-CONSTRAINED URBAN MODELS.

In the last chapter, the typology of models based on pseudo-dynamic forms gave rise to particular types of model in which the mover sequences associated with redistributing existing activity could be clearly used to enable locational constraints to be satisfied. From that discussion it emerged that the model type most suitable for this problem of constraint appeared to be that in which *all* activity could be reallocated. In this chapter, the way in which this type of model, referred to in Chapter 4 as an $\underline{\alpha}=\underline{I}$ type model, or more generally as a Baxter-Williams type model (Baxter and Williams, 1975), might be elaborated to handle locational constraints, will be described. An *ad hoc* constraints mechanism based on identifying appropriate values for $\underline{\alpha}$ will also be presented and finally applications will be made using data from the Reading subregion.

This chapter will pick up directly from the formal presentation of the last chapter and no new notation will be introduced. In fact, it is not possible to read this chapter in isolation from the last as the initial discussion in this chapter will concern the elaboration of $\underline{\alpha}=\underline{I}$ type models in terms of mechanisms used to redistribute the complete

pattern of activity or part of the sequence of activity generated in such multiplier models. Accordingly, the notation used and the equation systems referred to are those of Chapter 4. To begin the treatment then, methods for elaborating $\underline{\alpha}=\underline{I}$ type models will now be introduced as control type problems.

CONTROL THROUGH COMPLETE REDISTRIBUTION: COMPLETE SEQUENCES.

Consider the situation in which the first locational constraint violation occurs in the time period [R:R-1]. Such an occurrence might be based on some element or elements of $\underline{e}(R)$ or $\underline{p}(R)$ or both exceeding their respective constraints, $\underline{c}^e(R)$ and $\underline{c}^p(R)$. Up to the particular time period in which the first violation occurs, the model is effectively an $\underline{\alpha}=\underline{0}$ type model, and after this period it becomes an $\underline{\alpha}=\underline{I}$ model. In the following exposition, it is assumed that after the first violation, the constraints are violated in every succeeding time period until the simulation ends at r=t+T+1. This enables general forms for the model to be derived: in practice, constraints may not be violated in every time period and thus the resulting model and its mover processes would be a mix of $\underline{\alpha}=\underline{0}$ and $\underline{\alpha}=\underline{I}$ type models. Such models are too specific to present although they would pose no difficulties in terms of computation. Given this context, there are two possible forms for complete redistribution: complete redistribution of the whole sequence from initial input to the final amount of new change, and complete redistribution of only part of the sequence from the point at which new changes involve constraint violation to the final new change. These will now be examined in turn.

Up to the time r=R, the activity generation and distribution is given by equation (4.8) or (4.10) for employment and analogous equations exist

for population. After the constraint is violated, from $r=R+1$ to $r=t+1$, new complete mover sequences work themselves out $r+T+1$ time periods after the first constraint violation, therefore it is necessary to divide the overall simulation into five significant time intervals: $r=t-T$ and $t-T+1 \leq r \leq R$ which have already been considered and do not need to be made explicit again here, and $R+1 \leq R+n \leq t$, $t+1 \leq r \leq R+T+1$ and $R+T+2 \leq r \leq t+T+1$. An immediate intuitive grasp of this division is illustrated in Figure 5.1 where it is clear that the way in which mover sequences are begun leads to a slightly more complex form of model than the original Baxter-Williams type.

For $R+1 \leq R+n \leq t$, the employment $\underline{e}(R+n)$, $1 \leq n \leq t-R$, is given by

$$\underline{e}(R+n) = \underline{b}\{\underline{I}+ \sum_{w=t-T+n}^{R+n} \prod_{\tau=t-T}^{w} \underline{A}(\tau)\underline{B}(\tau)+ \sum_{v=R+1}^{R+n} \prod_{\tau=v}^{R+n} \underline{A}(\tau)\underline{B}(\tau)\}. \qquad (5.1)$$

Note here that the terms in original distribution which still affect the structure of employment are constant in number but change in form as the ones generated earlier are redistributed by the second term. Then from $t+1 \leq r \leq R+T+1$, the equation is made up of three major elements relating to the original distribution not yet moving, the activity stillbeing redistributed and the stayers associated with the original movement which has now ceased

$$\underline{e}(r) = \underline{b}\ \{\underline{I}+ \sum_{w=t-T+r-R}^{t} \prod_{\tau=t-T}^{w} \underline{A}(\tau)\underline{B}(\tau)+ \sum_{v=r+1}^{t+1} \prod_{\tau=v}^{r} \underline{A}(\tau)\ \underline{B}(\tau) \ +$$

$$\sum_{w=t-T}^{r-T-2} \prod_{z=t+1}^{w+T+1} \underline{A}(z)\underline{B}(z)\}. \qquad (5.2)$$

In Figure 5.1, the stippled boxes are associated with these elements and

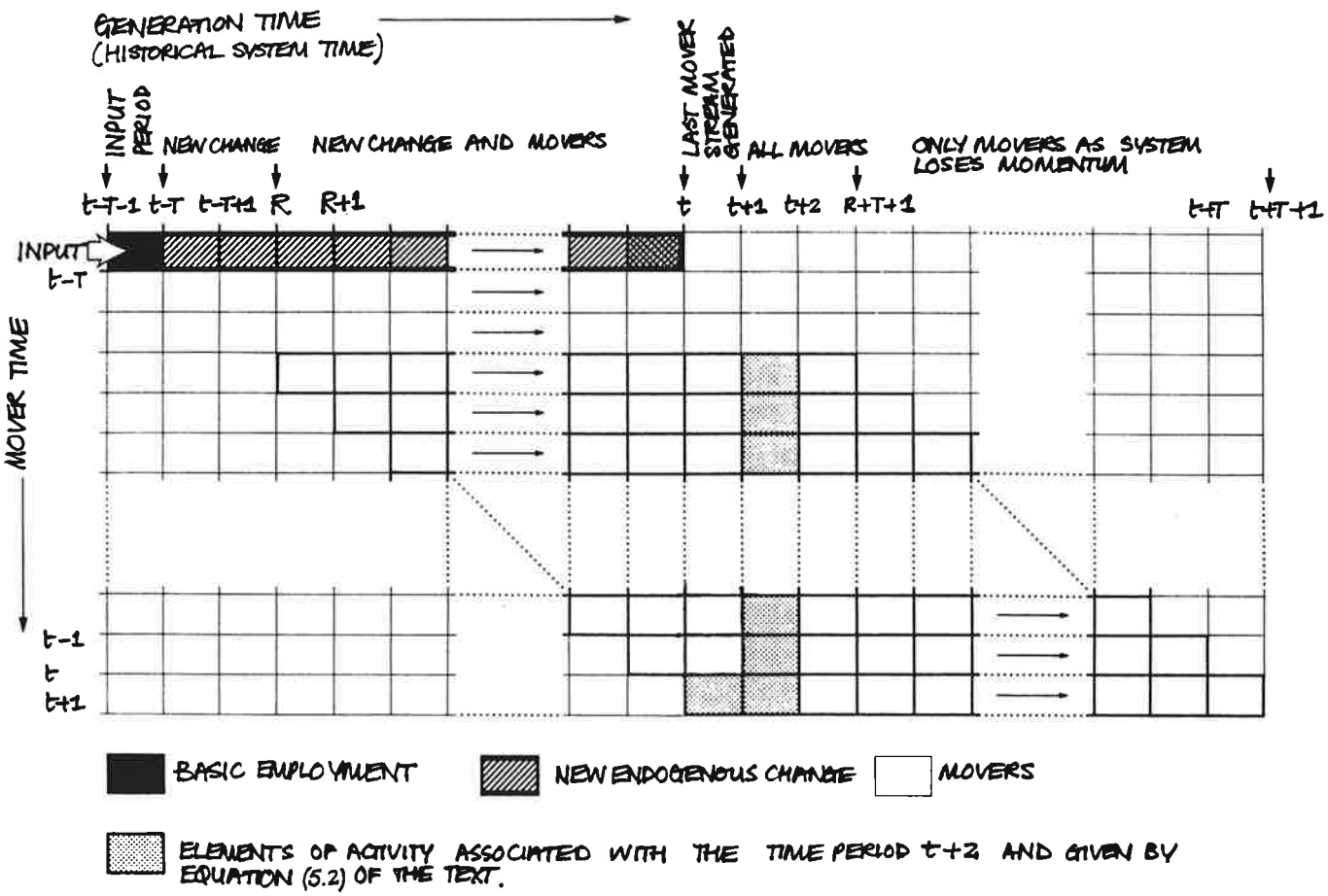



Figure 5.1: Mover Processes based on Redistribution of the Total Sequence of Change.

it is clear that they reflect different processes. In the final period, the employment is predicted from an appropriate modification of equation (4.14).

In this model, it is clear from Figure 5.1 and from equations (5.1) and (5.2) that only t+1-R mover processes are set in train due to the fact that the constraints only begin to be accounted for from $r=R+1$. If it is required to initiate T+1 mover sequences as in the original Baxter-Williams model, then the whole process would be pushed forward in time and a further equation would be required to show the activity predicted from $t \leq r \leq R+T+1$. This is easy to accomplish but adds nothing to the argument and thus it is excluded. There are many such variations which might be suitable, and every situation may demand its own variant. The framework developed here is sufficiently general to enable such variations to be made.

To meet the constraints assumed to be violated at every time $r \geq R$, it is necessary to find new distribution-redistribution matrices $\underline{A}(r)$ and $\underline{B}(r)$ which lead to $\underline{e}(r) \leq \underline{c}^e(r)$ and $\underline{p}(r) \leq \underline{c}^p(r)$. Traditionally, this problem has been solved in such models by applying row and column factors to the matrices based on the degree to which the constraints have not been met, and iterating until convergence. For the completely constrained matrix problem, convergence is assured and this has been demonstrated in a variety of fields: for example, in input-output analysis by Bacharach (1970) and in spatial interaction modelling by Evans (1970). For partially constrained problems, such as these presented here, such convergence to meet constraints cannot be proved but experience suggests that most problems studied so far are well behaved in that such procedures usually 'work'.

The problem here is more complicated in that the process of modifying $\underline{A}(r)$ and $\underline{B}(r)$ must operate through time, that is, no iteration is assumed to take place to find $\underline{A}(r)$ and $\underline{B}(r)$ at any time r. Then constraint violations are affected by the presence of new change up to r=t as well as by the process of trial and error adjustment of the matrices which will inevitably involve further constraint violations. However after time r=t, the process will continue without new change, and from this point, the emphasis in the model will be on redistributing the existing activity in an effort to meet the constraints. Two characteristics of the process are worthy of note: because of the process required to modify the matrices, convergence can never be assured, but it may be necessary to continue the simulation of whole sequences of activity after r=t+1, thus extending the simulation beyond t+T+1 so that the constraints can be satisfied. This would be a matter for experiment, for the life of the mover process T does not necessarily match the time required for this process to meet all the constraints.

To illustrate the process, assume that new matrices $\underline{A}(R+1)$ and $\underline{B}(R+1)$ are required so that the constraints on employment and population are met. It is possible to operate directly on the previous matrices $\underline{A}(R)$ and $\underline{B}(R)$ by applying factors based on the mismatch between the constraints and predicted activities. Such factors may be expressed generally by matrices $\underline{F}(R+1)$ and $\underline{G}(R+1)$ which reflect the positive feedback effects necessary to keep the system on course.

Then

$$\underline{A}(R+1) = \underline{A}(R)\underline{F}(R+1), \quad \text{and}$$

$$\underline{B}(R+1) = \underline{B}(R)\underline{G}(R+1).$$

If $\underline{F}(R+1)$ and $\underline{G}(R+1)$ are required to give $\underline{A}(R+1)$ and $\underline{B}(R+1)$ which ensure that $\underline{e}(R+1)$ and $\underline{p}(R+1)$ do not infringe any of their respective constraints, then it is likely that these factor matrices will have to be chosen by direct iteration within the time period [R+1:R]. However as suggested above, this is not assumed here but note that this does not imply that the constraints might never be reached. In this particular version, the nature of the choice for $\underline{A}(R+1)$ and $\underline{B}(R+1)$ does not assure that the constraints need be met by the re-distribution processes, although if required, such constraints can always be met. The process outlined here will probably converge and is computationally more efficient than those that are known to converge.

In many applications of the pseudo-dynamic model, it might be assumed that the scale effects of generation are independent of time in contrast to the distribution effects which are time dependent. As in previous chapters, we will assume a separability condition such that $\underline{A}(r) = \underline{\Lambda}\ \underline{T}(r)$ and $\underline{B}(r) = \underline{\Gamma}\ \underline{S}(r)$ where $\underline{\Lambda}$ and $\underline{\Gamma}$ are the respective scale effects based on diagonal scalar matrices and $\underline{T}(r)$ and $\underline{S}(r)$ are the distribution effects based on singly-stochastic (Markov) matrices. Using these assumptions, it is clear that

$$\underline{T}(R+1) = \underline{T}(R)\underline{F}(R+1) \quad \text{and}$$

$$\underline{S}(R+1) = \underline{S}(R)\underline{G}(R+1) \quad ,$$

and in general

$$\underline{T}(r) = \underline{T}(R) \prod_{\tau=R+1}^{r} \underline{F}(\tau) \quad \text{and}$$

$$\underline{S}(r) = \underline{S}(R) \prod_{\tau=R+1}^{r} \underline{G}(\tau) \quad .$$

In these terms, the problem of satisfying constraints is one of finding stable matrices $\underline{T}(r)=\underline{T}$ and $\underline{S}(r)=\underline{S}$.

At time r when these matrices have been found, it might then be necessary to continue the simulation to find a stable distribution of activity based on these forms. Then if $r \leq R+T+1$ where R indicates the time period when the first constraint is violated, then it can be deduced from equations (5.1) and (5.2) above that the final configuration of activity will be given by

$$\underline{e}(t+T+1) = \underline{b}[\underline{I} - \underline{T}\,\underline{\Lambda}\,\underline{\Gamma}\,\underline{S}]^{-1}.$$

This method is of particular importance in making operational Berechman's (1976) model, and in extending Baxter and Williams' (1975) model to deal with locational constraints. The method itself has only been sketched here and it will be presented in Chapters 8 and 9.

CONTROL THROUGH COMPLETE REDISTRIBUTION: PART SEQUENCES.

The second major process based on complete redistribution of activity in order to satisfy constraints, involves regenerating only part of the sequence of new change. In essence, the method involves beginning the process of regeneration at the point when a particular increment of new change leads to a constraint violation and only regenerating and redistributing that and succeeding increments of new change. The pattern of activities up to the time of the constraint violation is assumed to be stable and unchanging and only the pattern afterwards is subject to redistribution. In fact, the form of the model is much simpler than the one presented in equations (5.1)

and (5.2) for activity existing prior to the constraint violation does not need to be handled whereas it does in the above model.

Figure 5.2 illustrates this process diagrammatically and comparison with Figure 5.1 is sufficient to establish its relatively simple structure. In devising the appropriate equation for $\underline{e}(R)$, where R indicates the time period [R:R-1]in which constraints are first violated, it is necessary to develop the model from r=R-2. Then

$$\underline{e}(R-2) = \underline{b}\{\underline{I} + \sum_{w=t-T}^{R-2} \prod_{\tau=t-T}^{w} \underline{A}(\tau)\, \underline{B}(\tau) , \tag{5.3}$$

$$\Delta^*\underline{s}(R-1) = \underline{b} \prod_{\tau=t-T}^{R-1} \underline{A}(\tau)\, \underline{B}(\tau), \text{ and} \tag{5.4}$$

$$\Delta^*\underline{s}(R) = \Delta^*\underline{s}(R-1)\underline{A}(R)\underline{B}(R). \tag{5.5}$$

The model in equations (5.3) to (5.5) is clearly an $\underline{\alpha}=\underline{0}$ type and using these forms, the appropriate expression for $\underline{e}(R)$ can be stated

$$\underline{e}(R) = \underline{e}(R-2) + \Delta^*\underline{s}(R-1)[\underline{I}+\underline{A}(R)\underline{B}(R)]. \tag{5.6}$$

As previously, assume that the constraints are violated at R due to the new change $\Delta^*\underline{s}(R)$. Clearly, this new change $\Delta^*\underline{s}(R)$ and every successive element of $\Delta^*\underline{s}(r)$, r>R must be regenerated and redistributed with new matrices $\underline{A}(r)$ and $\underline{B}(r)$ based on positive feedback from the level of constraint violation. Then at R+1

$$\underline{e}(R+1) = \underline{e}(R-2)+ \Delta^*\underline{s}(R-1)\{\underline{I}+[\underline{I}+\underline{A}(R)\underline{B}(R)]\underline{A}(R+1)\underline{B}(R+1)\} , \tag{5.7}$$

and in general

$$\underline{e}(R+n) = \underline{e}(R-2) + \Delta^*\underline{s}(R-1)[\underline{I}+ \sum_{v=r}^{r+n} \prod_{\tau=v}^{R+n} \underline{A}(\tau)\underline{B}(\tau)]. \tag{5.8}$$

Equation (5.8) illustrates the essential structure characterising

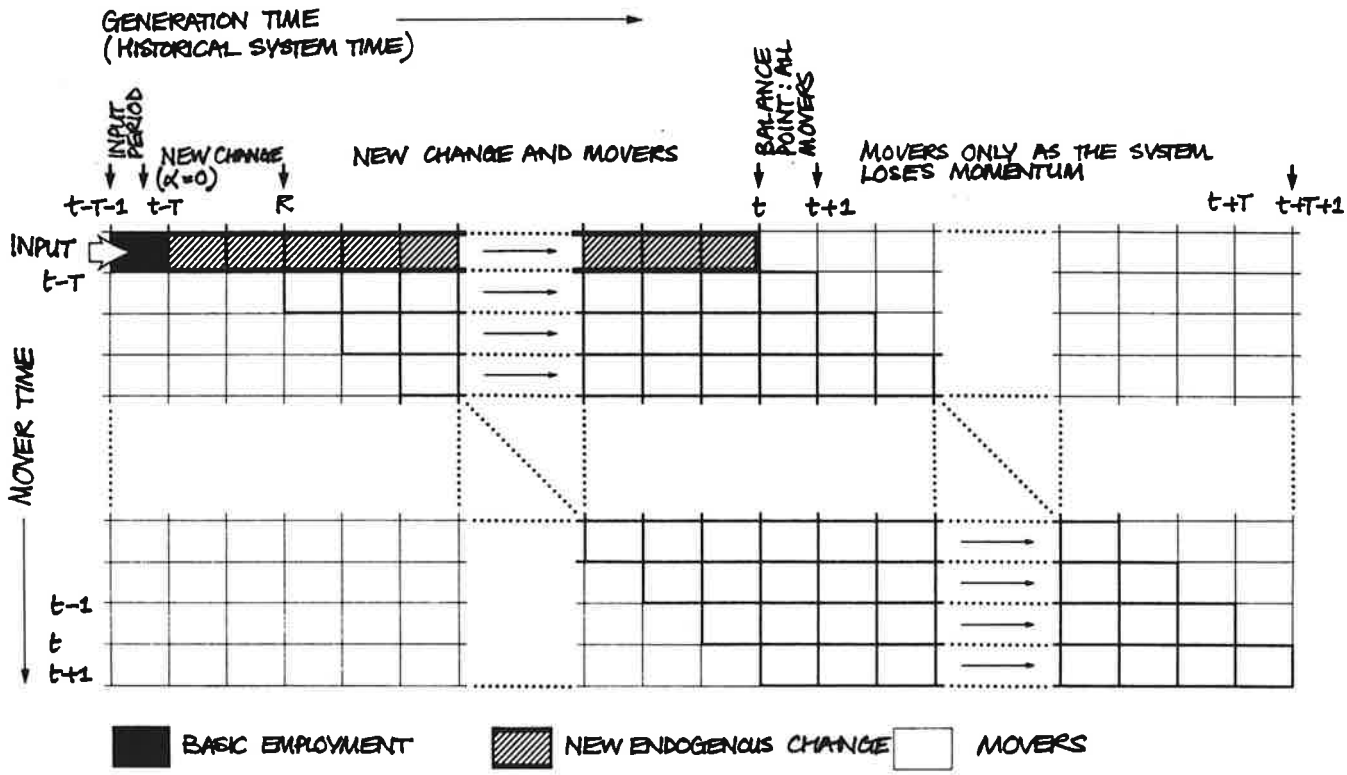



Figure 5.2: Mover Processes based on Redistribution of Part of the Sequence of Change.

the complete redistribution of part of the sequence of generation.

It is clear that $\underline{e}(R-1) = \underline{e}(R-2)+\Delta^*\underline{s}(R-1)$ remains stable as generated by the $\underline{\alpha}=\underline{0}$ model, and that the sequence starting with $\Delta^*\underline{s}(R)$ is regenerated according to the Baxter-Williams version of the $\underline{\alpha}=\underline{I}$ type model. It is as if $\Delta^*\underline{s}(R-1)$ is treated as the initial input $\underline{b}$ and the Baxter-Williams model applied thereafter. With regard to the form of the matrices $\underline{A}(r)$ and $\underline{B}(r)$, $r>R$, the above discussion in the previous section concerning the constrained matrix problem applies completely: the same caveats with regard to convergence are necessary, and the process of simulation may need to be longer than $t+T+1$ to attain a convergence which meets the constraints.

One final point remains to be made: if the constraints are only violated at irregular intervals, the reallocation may only occur with the same irregularity. That is, the partial mover sequences may only be initiated irregularly and thus the composite pattern would be somewhat different from that in equation (5.8). Specific forms could easily be worked out but as they lack the generality of this presentation, they have been omitted. In fact, it would be more appropriate to develop such forms for the second major type of constraint mechanism based on partial redistribution which is to be outlined below.

CONTROL THROUGH PARTIAL REDISTRIBUTION.

The basic idea characterising partial redistribution of existing

activity involves evaluating by how much any particular constraint or set of constraints is violated in any period [R:R-1], and initiating a mover sequence which regenerates and redistributes only the proportion of activity which violates the constraint. The proportion to be redistributed is set equal to the amount of activity moving which is controlled by the mover matrix $\underline{\alpha}$. Therefore, such a process of meeting constraints involves not only determining the matrices $\underline{A}(r)$ and $\underline{B}(r)$ but also determining the mover ratio matrix $\underline{\alpha}(\tau,z)$ where $\underline{\alpha}(\tau,z)$ is related to the time when the constraint violation occurs in the sense suggested below.

In the previous development of constraint procedures, general equations for the process were derived such as those in equations (5.1) and (5.2): here this is not really worthwhile as the original equations for the pseudo-dynamic model given previously as equations (4.1) to (4.7) are quite similar to those characterising the methods of this section. Indeed in the computable form for the model developed in the following section, a method of constraint based on the framework outlined here is used and specific equations are presented there. Note however that the following assumptions characterise the process: the constraints are first violated in time period [R:R-1] and are violated in every time period thereafter; the mover ratio matrix $\underline{\alpha}(\tau,z)$ is made time dependent in terms of z, not $\tau$, thus implying that each constraint violation initiates a mover sequence whose amount depends on the time of initiation z or constraint violation time z-1; and $\underline{A}(\tau)=\underline{\tilde{A}}(\tau)$ and $\underline{B}(\tau)=\underline{\tilde{B}}(\tau)$ which is the assumption of the Baxter-Williams model. Because $\underline{\alpha}$ is variable in the time dependent sense, none of the three previous

constant $\underline{\alpha}$ type models can be used to present a general form for the equation system.

The first model to be developed involves redistribution of the total sequence of new changes, the second redistribution of only part of the sequence, both in the manner suggested in the previous two sections. The procedure for fixing the value of $\underline{\alpha}(r)$, $\underline{A}(r)$ and $\underline{B}(r)$ is presented first for the method involving redistribution of the total sequence. Then in the time period [R:R-1], the following algorithm is used if constraints are violated: for the total employment in zone k at time R, a series of tests are made. If

$$E_k(R) \geq C_k^e(R), \tag{5.9}$$

zone k is assigned to the set of constrained employment zones $Z_e$ and the surplus employment $\Delta_k^e(R)$ is calculated from

$$\Delta_k^e(R) = E_k(R) - C_k^e(R), \quad k\varepsilon Z_e. \tag{5.10}$$

As a proportion of the total employment to be generated, $\Sigma_k E_k$, the surplus is expressed as the ratio $\sigma(R)$ computed as

$$\sigma(R) = \sum_{k\varepsilon Z_e} \Delta_k^e(R)/\sum_k E_k \quad . \tag{5.11}$$

If the constraint on employment has been violated, it is then necessary to normalise the previous distribution matrix $\underline{B}(R)$ so that no further activity is allocated to the constrained zones. Then

$$B_{jk}(R+1) = 0, \quad k\varepsilon Z_e, \forall_j \ , \tag{5.12}$$

and the $\underline{B}(R+1)$ matrix is structured to ensure that the matrix $\underline{S}(R+1)$ is row stochastic.

The same process is used to check the population constraints. If

$$P_j(R) \geq C_j^p(R), \tag{5.13}$$

zone j is assigned to set $Z_p$, and the surplus $\Delta_j^p(R)$ is calculated

$$\Delta_j^p(R) = P_j(R) - C_j^p(R), \qquad j\varepsilon Z_p \quad . \tag{5.14}$$

The ratio $\rho(R)$ is now computed from

$$\rho(R) = \sum_{j\varepsilon Z_p} \Delta_j^p(R)/\sum_j P_j \ , \tag{5.15}$$

and the $\underline{A}(R)$ matrix is normalised to account for constraint violations. Then

$$A_{ij}(R+1) = 0, \quad j\varepsilon Z_p, \forall_i \quad , \tag{5.16}$$

where $\underline{A}(R+1)$ is structured to ensure that $\underline{T}(R+1)$ is row stochastic.

Because of the interdependence of employment and population, it is necessary to define $\alpha_{ii}(R+1)$ as

$$\alpha_{ii}(R+1) = \max[\sigma(R),\rho(R)], \ \forall_i \ , \tag{5.17}$$

which will make the model redistribute all activity so that the most severe constraint violation dominates the process. The pseudo-dynamic model is then operated in the normal fashion noting that $\alpha_{ii}(R+1)$ pertains to the appropriate stream of activity resulting from constraint violation at R. Therefore at any one point in time r>R, there are several streams being reallocated, each according to $\underline{\alpha}(\tau)$, $R+1 \leq \tau \leq r$.

In redistributing the total sequence of activity, it is argued that any constraint violation depends not upon a specific problem in any

particular zone, but on a systematic distortion of the whole process of distribution. Therefore to correct such system-wide failures in the original distribution, it is assumed that the proportion to be redistributed applies to the total of existing activity which can be easily calculated from the initial input $\underline{b}$ and the known matrices $\underline{\Lambda}$ and $\underline{\Gamma}$. That is the total employment $\Sigma_k E_k$ can be calculated from $\underline{b}[\underline{I}-\underline{\Lambda}\ \underline{\Gamma}]^{-1}$ which is equal to $\Sigma_k b_k/(1-\lambda\gamma)$ when $\underline{\Lambda}$ and $\underline{\Gamma}$ are scalar diagonal, and the population can be calculated from $\lambda\Sigma_k E_k$ in the same case.

The second major case which involves a partial sequence of movers, depends upon the redistribution of the new change which leads to violation of the constraint, and further changes generated in the same sequence. A system of equations similar to (5.9) to (5.17) is used for dealing with the constraints except that $\Sigma_k E_k$ and $\Sigma_j P_j$ in equations (5.11) and (5.15) respectively, are replaced with the sums appropriate to the reallocation of the rest of the sequence, not the whole sequence. In these particular instances, these methods of constraint based on partial redistribution are really only suitable if $\underline{\Lambda}$ and $\underline{\Gamma}$ are scalar diagonal. However it is possible to consider a variety of other methods of partial redistribution in which the ratios $\sigma(R)$ and $\rho(R)$ refer to various parts of the sequence of change generated so far; and in such cases, there is no necessity for calculating $\Sigma_k E_k$ and $\Sigma_j P_j$ from other formulas, for these relate to what has been already generated.

As in the method of constraint developed in the previous section,

considerable variation exists within this general framework, as will be illustrated in the example next described, and the ones presented in Chapters 6 to 9 of this thesis. It is also possible to make $\sigma_k(R)$ and $\rho_j(R)$ zone specific, thus embracing the kind of constraint processes developed by Echenique, Crowther and Lindsay (1969). As this would make the form of pseudo-dynamic model slightly different in terms of equation system from the one presented above, it is not pursued further here. It will, however, be taken up again in the next chapter where a zone specific set of ratios is developed for the treatment of constraints.

## A COMPUTABLE FORM FOR A PSEUDO-DYNAMIC ACTIVITY ALLOCATION MODEL.

A central argument in support of the pseudo-dynamic models introduced here rests on the notion that by making the implied dynamic processes within static models more explicit, major advantages concerning substantive interpretation and operational tractability will result. To demonstrate this point, it is now proposed to apply the theoretical developments of this and the previous chapters to conventional forms of activity allocation model. By way of conclusion, a model which incorporates one of the methods of constraint previously developed and utilises the information-minimising framework presented in Chapter 3 will be outlined, thus showing how constraints on the artificial growth of an urban area can be handled using very simple inputs and assumptions. The information-minimising method will be used to build up the distribution and redistribution matrices $\underline{A}(\tau)$ and $\underline{B}(\tau)$ from acceptable prior information about the system and this will also show how much prior information is required for such model structures. In

the next chapter this logic will be taken further when a calibration method is developed using the model's dynamic processes.

The model will be presented here in a more conventional form which illustrates the structure used for computation. In this sense, the iterative-recursive form on which the original dynamic model was developed in Chapter 3 is relevant, and the form used here mirrors conventional activity allocation models in current usage (Batty, 1976). Two assumptions must be clarified: the constraint process used is based on the method of partial redistribution outlined in the previous section, and the regeneration and redistribution of activity which is involved in constraint violations is based on the complete sequence of new change. Both population and employment equations will be stated, thus showing their explicit interdependence. It is also assumed that $\underline{A}(\tau) = \underline{\tilde{A}}(\tau)$ and $\underline{B}(\tau) = \underline{\tilde{B}}(\tau)$ as in Baxter-Williams type models.

The computable form of the model is built up in stages: first new change, then movers, finally stayers for population and employment respectively. New population change $\Delta^*\underline{p}(r)$ and employment change $\Delta^*\underline{s}(r)$ are given as

$$\Delta^*\underline{p}(r) = \Delta^*\underline{s}(r-1)\underline{A}(r), \tag{5.18}$$

$$\Delta^*\underline{s}(r) = \Delta^*\underline{p}(r)\underline{B}(r) = \Delta^*\underline{s}(r-1)\underline{A}(r)\underline{B}(r). \tag{5.19}$$

For population movers $\underline{p}^m(r,w)$ and employment movers $\underline{s}^m(r,w)$, the relevant recurrence relations are

$$\left.\begin{aligned} \underline{p}^m(r,w) &= \underline{s}^m(r-1,w-1)\underline{A}(r), \quad w>t-T, \\ \underline{p}^m(r,t-T) &= \Delta^*\underline{s}(0)\underline{\alpha}(r)\underline{A}(r) , \end{aligned}\right\} \tag{5.20}$$

$$\left.\begin{aligned} \underline{s}^m(r,w) &= \underline{p}^m(r,w)\underline{B}(r) = \underline{s}^m(r-1,w-1)\underline{A}(r)\underline{B}(r), \quad w>t-T \\ \underline{s}^m(r,t-T) &= \Delta^*\underline{s}(0)\underline{\alpha}(r)\underline{A}(r)\underline{B}(r). \end{aligned}\right\} \quad (5.21)$$

For population stayers $\underline{p}^s(r,w)$ and employment stayers $\underline{s}^s(r,w)$, the equations are a little simpler in that no distribution matrices are involved directly

$$\underline{p}^s(r,w) = [\underline{p}^m(r-1,w) + \underline{p}^s(r-1,w)][\underline{I}-\underline{\alpha}(r-w+1)], \quad (5.22)$$

$$\underline{s}^s(r,w) = [\underline{s}^m(r-1,w)+ \underline{s}^s(r-1,w)][\underline{I}-\underline{\alpha}(r-w+1)]. \quad (5.23)$$

Equations (5.18) to (5.23) indicate that the process is dependent on certain initial conditions for the initial input employment $\Delta^*\underline{s}(0)$, and the movers and stayers. These conditions can be listed as follows:

$$\Delta^*\underline{s}(0) = \underline{b}\ , \quad (5.24)$$

$$\underline{s}^m(r-1,r-1) = \underline{0};\ \ \underline{p}^m(r-1,r-1) = \underline{0}, \quad \text{and} \quad (5.25)$$

$$\underline{s}^s(r-1,r-1) = \Delta^*\underline{s}(r-1);\ \ \underline{p}^s(r-1,r-1) = \Delta^*\underline{p}(r-1)\ . \quad (5.26)$$

Because the time scale for such a model is clearly not historical time, although it might be regarded as some approximation to this, it will be assumed that the time index r relates to computer iteration time; thus when r = 0, this indicates the start of the process symbolising existence of the input, so that $\Delta\underline{s}^*(0) = \underline{b}$ as in equation (5.24) above, and if $\underline{A}(r)$ and $\underline{B}(r)$ are functions of previous values of the same matrices, then $\underline{A}(0)$ and $\underline{B}(0)$ represent prior distributions known before the simulation begins. Thus the process which runs from $t-T\leq r\leq t+T+1$, now specifically runs from $1\leq r\leq 2T+2$, which is a neat indication that the life of the simulation is twice the life of a single generation-regeneration sequence.

It is only necessary to develop the model for the employment equations as the population equations follow directly. Then for $1<r\leq T+2$,

$$\underline{e}(r) = \Delta^*\underline{s}(0)+\Delta^*\underline{s}(r)+\sum_{w=1}^{r-1}\underline{s}^m(r,w) + \sum_{w=1}^{r-1}\underline{s}^s(r,w),$$

$$= \Delta^*\underline{s}(0)+[\Delta^*\underline{s}(r-1) + \sum_{w=2}^{r-1}\underline{s}^m(r-1,w-1)+\Delta^*\underline{s}(0)\underline{\alpha}(r)]\underline{A}(r)\underline{B}(r) +$$

$$\sum_{w=1}^{r-1}[\underline{s}^m(r-1,w)+\underline{s}^s(r-1,w)][\underline{I}-\underline{\alpha}(r-w+1)]. \qquad (5.27)$$

Note that when $r=T+2$, $\Delta^*\underline{s}(T+2)=\underline{0}$. Although equation (5.27) is specific to the interval $1<r\leq T+2$, it could easily be used for any time period in the whole simulation, simply noting the initial conditions and the fact that $\underline{\alpha}(\tau)=\underline{0}$, $\tau>T+2$.

In fact, in computing the model, storage space is reserved in the program for variables for each period of time from 1 to 2T+2; although this is certainly not the most economical organisation, it makes the programming much easier, and it is essential when a flexible program in which T may vary from application to application, is required.

For the time period $T+3\leq r\leq 2T+2$, $\underline{e}(r)$ is computed as

$$\underline{e}(r) = \Delta^*\underline{s}(0) + \sum_{w=r-T-1}^{T+1}\underline{s}^m(r,w) + \sum_{w=r-T-1}^{T+1}\underline{s}^s(r,w)$$

$$+ \sum_{w=1}^{r-T-2}\underline{s}^s(w+T+1,w) ,$$

$$= \Delta^*\underline{s}(0) + \sum_{w=r-T-1}^{T+1}\underline{s}^m(r-1,w-1)\underline{A}(r)\underline{B}(r)$$

$$+ \sum_{w=r-T-1}^{T+1}[\underline{s}^s(r-1,w)+\underline{s}^m(r-1,w)][\underline{I}-\underline{\alpha}(r-w+1)]$$

$$+ \sum_{w=1}^{r-T-2}\underline{s}^s(w+T+1,w). \qquad (5.28)$$

Equations (5.27) to (5.28) are both functions of the previous state of the system which derives from the initial conditions, as well as the distribution matrices $\underline{A}(r)$ and $\underline{B}(r)$, and the mover ratio matrix $\underline{\alpha}(r)$. The distribution and mover matrices may be exogenous but in this instance, they are based on two subprocesses: the process involving constraint violation, and the process involving lagged relationships between distribution from one time period to the next.

The method of fixing $\underline{\alpha}(r)$ and normalising $\underline{A}(r)$ and $\underline{B}(r)$ to ensure no further violations occur, has already been described in the previous section. At the end of each time period, the constraints are checked and equations (5.9) to (5.17) are solved if necessary. If no constraints are violated, then $\underline{\alpha}(r)=\underline{0}$ and $\underline{A}(r)$ and $\underline{B}(r)$ do not need to be renormalised. To complete the presentation of this model, it is now necessary to derive explicit forms for the distribution matrices which relate to the notion of lags in distribution and initial prior information about the pattern of location, before applications of the model are described.

DYNAMIC FORMS FOR SPATIAL INTERACTION-DISTRIBUTION.

It is assumed that the scale effect within the distribution matrices $\underline{A}(r)$ and $\underline{B}(r)$ of the pseudo-dynamic model is independent of time, and constant across space. Then formally

$$\underline{A}(r) = \underline{T}(r)\underline{\Lambda}, \quad \text{and} \tag{5.29}$$

$$\underline{B}(r) = \underline{\Gamma}\ \underline{S}(r), \tag{5.30}$$

as was assumed in the previous discussion concerning locational constraints. The scalar diagonal matrices $\underline{\Lambda}$ and $\underline{\Gamma}$ must be known before

the simulation begins and this is usual if total employment and population can be observed or predicted independently. Therefore, it is the matrices $\underline{T}(r)$ and $\underline{S}(r)$ which need to be explored here.

An obvious constraint on their form requires that they be row stochastic, so that they act as true distribution matrices, allocating employment and population respectively. Then if $t_{ij}(r)$ and $s_{jk}(r)$ indicate the respective elements of these matrices, it is necessary that

$$\sum_j t_{ij}(r) = 1, \quad \text{and} \quad \sum_k s_{jk}(r) = 1.$$

Note that the subscript j refers to residential (population) zones, and i and k to workplace (employment) zones. It is possible to model $t_{ij}(r)$ and $s_{jk}(r)$ using data which is completely exogenous to the general model, but it is likely that these distribution models will depend to a certain extent upon previous distributions, that is, contain certain autoregressive terms. Furthermore, it is possible that distribution patterns in two sectors, say the population-service sectors, will influence distribution patterns in another two sectors, say employment-population.

These two types of interrelationship - through time and between sectors - have in fact been adopted in the specification of the T(r) and $\underline{S}(r)$ matrices for the model applied here. Many other hypotheses are possible, but it was felt that these notions would illustrate the potential of the idea, and moreover, if such dependence does not exist, the calibration would account for this. It was hypothesised that the pattern of distribution from employment to population - the journey-to-work -

could be derived from the previous distribution linking population to services by the application of new information concerning the difference between them, encoded in the matrix $\underline{F}(r)$. Then

$$\underline{T}(r) = \underline{S}(r-1)\underline{F}(r), \tag{5.31}$$

and the same type of relationship in which $\underline{S}(r)$ could be derived from $\underline{T}(r)$ by the application of information matrix $\underline{G}(r)$ was postulated

$$\begin{aligned} \underline{S}(r) &= \underline{T}(r)\underline{G}(r), \\ &= \underline{S}(r-1)\underline{F}(r)\underline{G}(r). \end{aligned} \tag{5.32}$$

Equation (5.32) establishes the basic recurrence relation. In a process which starts with r=1, it is necessary to have a prior distribution (r=0) which begins the process of successive distribution.

Looking at the structure implied by equations (5.31) and (5.32) it is clear that the first equation is

$$\underline{T}(1) = \underline{S}(0)\underline{G}(1),$$

and therefore the prior distribution matrix $\underline{S}(0)$ must be known before the simulation starts. In essence, the process depends upon an initial distribution which is successively modified by the information matrices $\underline{F}(r)$ and $\underline{G}(r)$; these incorporate new exogenous information, or indeed endogenous information which is being generated from the state of the system and used in a positive feedback sense. This clearly indicates a connection to the ways in which locational constraints are met which was presented above, although here, the constraint procedure acts over these lagged equations. The initial distribution matrix $\underline{S}(0)$ reflects prior knowledge which affects location, and might be based on a specification of the physical effects of space, in the sense suggested previously by the author (Batty and March, 1976).

Using the recurrence relations established in equations (5.31) and (5.32), it is possible to show how any pattern of distribution between two sectors is formed from the initial prior distribution matrix $\underline{S}(0)$ by the successive application of new information encoded in the matrices $\underline{F}(\tau)$ and $\underline{G}(\tau)$. Assuming the iterative sequence $1 \leq \tau \leq R$, forms for $\underline{A}(R)$ and $\underline{B}(R)$ are derived as follows:

$$\underline{A}(R) = \underline{T}(R)\underline{\Lambda} = \underline{S}(R-1)\underline{F}(R)\underline{\Lambda} ,$$

$$= \underline{S}(0) \{ \prod_{\tau=1}^{R-1} \underline{F}(\tau)\underline{G}(\tau)\}\underline{F}(R)\underline{\Lambda} \qquad (5.33)$$

$$\underline{B}(R) = \underline{\Gamma}\ \underline{S}(R) = \underline{\Gamma}\ \underline{S}(R-1)\underline{F}(R)\underline{G}(R) ,$$

$$= \underline{\Gamma S}(0) \{ \prod_{\tau=1}^{R} \underline{F}(\tau)\underline{G}(\tau)\}. \qquad (5.34)$$

At each iteration R, the matrices $\underline{A}(R)$ and $\underline{B}(R)$ must be estimated from new information which is supplied to the system exogenously, for example, in terms of some constraint, or endogenously from the current state of the system. A consistent means of generating information matrices $\underline{A}(\tau)$ and $\underline{B}(\tau)$ is through an information-minimising scheme similar to that outlined in Chapter 3. Such a scheme appropriate to the hypotheses suggested in the above discussion, is fairly easy to present and leads to conventional forms for the various submodels.

DISTRIBUTION MATRICES BASED ON INFORMATION-MINIMISING.

To estimate forms for $\underline{T}(r)$ and $\underline{S}(r)$, two associated probability distributions must be defined. First define the distribution $p_{ij}(r)$, $\Sigma_i\Sigma_j p_{ij}(r)=1$ which relates to the probability $t_{ij}(r)$, and then define $q_{jk}(r)$, $\Sigma_j\Sigma_k q_{jk}(r)=1$, relating to $s_{jk}(r)$. Following the discussion of information-minimising in Chapter 3, a first order minimisation is suggested in which the two

sectors are related by a pattern of distribution reflecting the interaction probabilities associated with both. Examining the employment-population (journey-to-work) distribution first, it is necessary to minimise an information function $I_1(r,r-1)$ subject to certain endogenous constraints on location and exogenous constraints on interaction. Then

$$\min I_1(r,r-1) = \min \sum_{ij}\sum p_{ij}(r) \ln \frac{p_{ij}(r)}{q_{jk}(r)} \quad , \quad i=k, \tag{5.35}$$

subject to

$$\sum_j p_{ij}(r) = \sum_j q_{jk}(r-1), \quad i=k, \quad \text{and} \tag{5.36}$$

$$\sum_{ij}\sum p_{ij}(r)c_{ij}(r-1) = \bar{C}(r) \ . \tag{5.37}$$

Note that equation (5.36) relates the location of employment used to allocate population at time r to the previous location of employment at r-1, and equation (5.37) reflects an exogenous constraint on the cost of travel: $c_{ij}(r-1)$ is the cost of travel from origin i to destination j lagged one time period to r-1 and $\bar{C}(r)$ is the mean cost of such travel.

Using the usual method of minimisation of a constrained function (see Webber, 1979) leads to a first order interaction model of the form

$$p_{ij}(r) = [\sum_j q_{jk}(r-1)] \frac{q_{jk}(r-1) \exp\{-\mu_1(r)c_{ij}(r-1)\}}{\sum_j q_{jk}(r-1) \exp\{-\mu_1(r)c_{ij}(r-1)\}} \quad , \ i=k \tag{5.38}$$

$\mu_1(r)$ is a parameter of the exponential function which must be calibrated at r so that constraint equation (5.37) is satisfied. The normalised probability $t_{ij}(r)$ is derived from equation (5.38) by

$$t_{ij}(r) = \frac{p_{ij}(r)}{\sum_j p_{ij}(r)} \quad , \tag{5.39}$$

and it is clear that the explicit form for equation (5.38) is the form for a conventional origin-constrained spatial interaction model.

The probability $s_{jk}(r)$ depends upon $t_{ij}(r)$ according to equation (5.32), and the appropriate information-minimising scheme involves minimisation of the first order function $I_2(r, r-1)$

$$\min I_2(r,r-1) = \min \sum_{jk}\sum q_{ij}(r) \ln \frac{q_{jk}(r)}{p_{ij}(r)} , \quad k=i , \tag{5.40}$$

subject to

$$\sum_k q_{jk}(r) = \sum_i p_{ij}(r), \; k=i, \quad \text{and} \tag{5.41}$$

$$\sum_{jk}\sum q_{jk}(r)c_{jk}(r-1) = \bar{S}(r). \tag{5.42}$$

Equation (5.41) is the endogenous origin constraint calculated from equation (5.38) and equation (5.42) is the constraint on average travel cost $\bar{S}(r)$. Minimisation leads to the form

$$q_{jk}(r) = [\sum_i p_{ij}(r)] \frac{p_{ij}(r) \exp\{-\mu_2(r)c_{jk}(r-1)\}}{\sum_{i,k} p_{ij}(r) \exp\{-\mu_2(r)c_{jk}(r-1)\}} , \quad k=i, \tag{5.43}$$

where $\mu_2(r)$ is the parameter controlling the average travel cost constraint $\bar{S}(r)$. The normalised probability $s_{jk}(r)$ is calculated as

$$s_{jk}(r) = \frac{q_{jk}(r)}{\sum_k q_{jk}(r)} , \tag{5.44}$$

and it is clear that the model has the same origin-constrained structure as the one above. The prior probability distribution $\{q_{jk}(0)\}$ will be specified in the next section where the application and calibration of the complete model is described.

A word of explanation concerning the subscripts used in equations (5.35) to (5.44) is necessary: both matrices, $\underline{T}(r)$ and $\underline{S}(r)$, are origin-constrained in the sense implied by equations (5.39) and (5.44). However, origins in the workplace-residential model are in terms of workplaces, in the residential-service centre model in terms of residences. Consequently the subscript j refers to residences, i and k to workplaces, thus the logical sequence $i \rightarrow j \rightarrow k$ reflects the multiplication of the distribution matrices $\underline{A}(r)\underline{B}(r)$. The different subscripting of $p_{ij}(r)$ and $q_{jk}(r)$ leads to difficulties when they come to be related until it is realised that the subscripts i and k refer to the same set of locations and are thus equivalent.

When the model is operated, the elements of the matrices $\underline{T}(r)$ and $\underline{S}(r)$ are computed using the non-matrix equations (5.38) and (5.39), (5.43) and (5.44) respectively. In fact, the matrices $\underline{F}(r)$ and $\underline{G}(r)$ have no simple form and are not computed as such, but they have been included here to illustrate the long term effect of the information-minimising process. It would be a simple matter to calculate them at each time r from

$$\underline{F}(r) = [\underline{S}(r-1)]^{-1}\underline{T}(r), \quad \text{and}$$

$$\underline{G}(r) = [\underline{T}(r)]^{-1}\underline{S}(r) \quad = [\underline{S}(r-1)\underline{F}(r)]^{-1}\underline{S}(r) \ ,$$

and this might yield useful information concerning the change in distribution through time. From both the matrix and non-matrix equations describing the effect of information change on the previous distributions, it is clear that when no new information is available, that is, when the previous distribution already meets the constraints, the information matrices $\underline{F}(r)=\underline{G}(r)=\underline{I}$, and the parameters of the submodels in equations (5.38) and (5.43) would equal zero. This would be most unlikely unless the constraint on travel cost were constant from iteration to iteration, and unless the process had converged to some kind of locational equilibrium.

One final point needs to be mentioned: the matrices $\underline{A}(\tau)$ and $\underline{B}(\tau)$ are controlled by the lagged process specified in equations (5.33) and (5.34) made explicit in this section, and by the constraint mechanism. In the computation of the model, the matrices are first calculated according to interaction submodels in equations (5.38) and (5.43), and after this, they are renormalised according to any violation of the constraints. Thus the constraint violation procedure takes precedence but note that once constraints have been violated and the matrices appropriately adjusted, their form is preserved by the equations used to compute the spatial interaction submodels.

CALIBRATION OF THE MODEL TO THE READING SUBREGION.

A further simplification was necessary in the application of the model whose form has been outlined in equations (5.18) to (5.44), and this relates to the variation in travel costs over time, in this case through the iterations. As it is impossible to assemble a meaningful time series of data in this regard, it is assumed that the travel costs and the associated parameters are independent of time r. Thus $\bar{C}(r)=\bar{C}$ and $\bar{S}(r) = \bar{S}$, and these values are observable for the whole process. In fact, this means that $\mu_1(r)=\mu_1$ and $\mu_2(r)=\mu_2$ and these parameters are chosen to satisfy the overall constraints

$$\sum_{ij}\sum p_{ij}c_{ij} = \bar{C} \text{ , and} \tag{5.45}$$

$$\sum_{jk}\sum q_{jk}c_{jk} = \bar{S} \text{ .} \tag{5.46}$$

Within this problem, the individual submodels which predict $p_{ij}(r)$ and $q_{jk}(r)$ now become

$$p_{ij}(r) = [\sum_j q_{jk}(r-1)] \frac{q_{jk}(r-1)\ \exp\{-\mu_1 c_{ij}\}}{\sum_j q_{jk}(r-1)\ \exp\{-\mu_1 c_{ij}\}}\ , \ i = k, \qquad (5.47)$$

and

$$q_{jk}(r) = [\sum_i p_{ij}(r-1)] \frac{p_{ij}(r)\ \exp\{-\mu_2 c_{jk}\}}{\sum_{i,k} p_{ij}(r)\ \exp\{-\mu_2 c_{jk}\}}\ , \ k=i. \qquad (5.48)$$

Note that in this case, the global probability distributions $\{p_{ij}\}$ and $\{q_{jk}\}$ which are derived from the final composite patterns of distribution are consistent with an information-minimisation relating to these composite patterns, and equations (5.47) and (5.48) although consistent with this structure, are only derivable in an *ad hoc* manner.

In essence, these simplifications mean that the calibration problem is no longer dynamic but static in that parameters $\mu_1$ and $\mu_2$ must be estimated for constraints which describe the average travel cost for the whole simulation period. Thus, it is possible to use a static method of estimation in which the parameters $\mu_1$ and $\mu_2$ are found by solving equations (5.45) and (5.46). Although the static calibration procedure does account for dynamic change in the pattern of distribution, it is the overall pattern of distribution which is the subject of the calibration due to the fact that only $\bar{C}$ and $\bar{S}$ are observable. Equations (5.45) and (5.46) involve the complete solution of the model in predicting $\{p_{ij}\}$ and $\{q_{jk}\}$ for given $\mu_1$ and $\mu_2$; clearly the equations are simultaneous and nonlinear, and their solution requires some numerical algorithm. In the example here, solutions were obtained using the Newton-Raphson method (Batty, 1976).

In the next chapter, the calibration problem will also be made dynamic in the sense implied by the information-minimisation of the previous section.

Because the static calibration problem itself can only be solved iteratively, the key to more efficient calibration involves matching the iterative solution procedure to the iterative structure of the pseudo-dynamic model. This is achieved in the next chapter in the same kind of way in which the iterative structure of the constraint procedures were dealt with here. The dynamic structure of the distribution process still remains without an interpretation of the process at a micro level in information-minimising terms. Thus a prior form for $\underline{S}(0)$ is still required.

In interaction models of this kind, it has been argued by the author (Batty and March, 1976) that an appropriate prior which should be explicitly accepted, is the Coleman-Zipf model based on the physics of the space. This model is in effect a two-dimensional gravity model which is based upon the peripheral and centralising influences contained in any bounded region. Its use in this model implies that the new information which modulates the prior concerns the behavioural characteristics of travel whereas the prior itself is based on strictly physical constraints on travel.

All the elements have now been presented which form the basis of the model applied to the Reading subregion. The model is designed to simulate the static configuration of activities in 18 zones using data taken from various surveys and censuses at 1966. In every sense, this example is a demonstration that the model is a useful way of building up an artificial growth process from rather simple input data. The small number of zones used is a characteristic of the hypothetical nature of this exercise, which is included here as much by way of a statement that the pseudo-dynamic model is an operational one, as for any substantive interpretations

which might be drawn about modelling the Reading subregion.

The zoning system used by the model is shown in Figure 5.3. Nevertheless, there are certain more substantive points to be made concerning the results of the model, largely because the model has been run using three different forms of prior distribution. The first two models use a prior relating to the dynamic scheme shown in equations (5.33) and (5.34) whereas the third model is based on two priors. Then in this model

$$\underline{T}(r) = \underline{T}(r-1)\underline{F}(r), \quad \text{and} \tag{5.49}$$

$$\underline{S}(r) = \underline{S}(r-1)\underline{G}(r). \tag{5.50}$$

Thus to start the process, priors based on $\underline{T}(0)$ and $\underline{S}(0)$ are required and this third model does not hypothesise any relationship between the sectors.

These three forms are presented in Table 5.1 and it is immediately clear that the first and third models use an extremely simple prior based on land and distance in the system. In effect, this means that the prior probability of interaction between any two zones is based solely on the dimensional qualities of the system. In the second model, the prior is based upon the known distribution of population at the base date as well as distance. Thus in the first and third models, there is no possibility of a tautology in which the model is predicting an activity on the basis of knowledge about that same activity, whereas in the second model, this is the case. Another less important difference between these three model forms relates to their autoregressive structure. In the third model $\underline{T}(r)$ and $\underline{S}(r)$ are functions of new information $\underline{F}(r)$ and

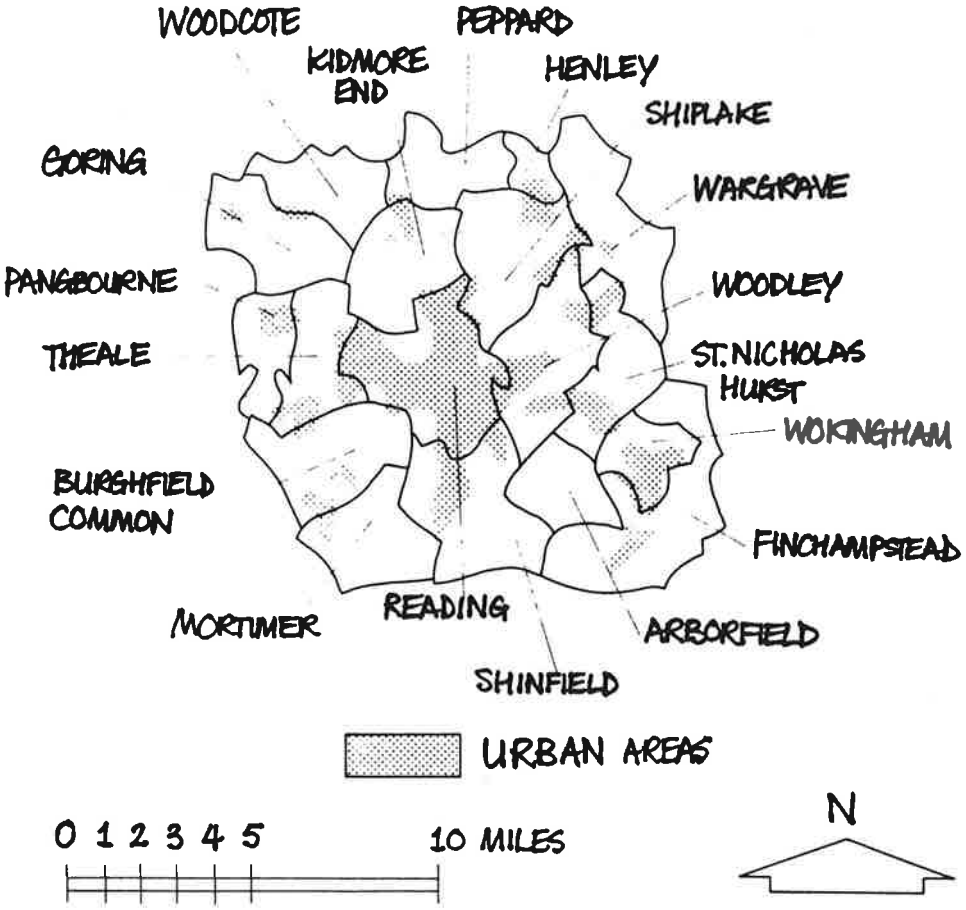



Figure 5.3: Zoning System for the Reading Model.

Table 5.1: Forms of Activity Allocation Model Based on Different Prior Distributions.

| Model | Initial Prior(s) on Iteration (0) | $\underline{T}(r)$ Matrix Function on Iteration (r) | $\underline{S}(r)$ Matrix Function on Iteration (r) |
|---|---|---|---|
| 1 | $s_{jk}(0) \propto L_j d_{jk}^{-1}$ | $\underline{T}(r) = \underline{S}(r-1)\underline{F}(r)$ | $\underline{S}(r) = \underline{T}(r)\underline{G}(r)$ |
| 2 | $s_{jk}(0) \propto P_j d_{jk}^{-1}$ | $\underline{T}(r) = \underline{S}(r-1)\underline{F}(r)$ | $\underline{S}(r) = \underline{T}(r)\underline{G}(r)$ |
| 3 | $\left\{ \begin{array}{l} t_{ij}(0) \propto L_j d_{ij}^{-1} \\ s_{jk}(0) \propto L_j d_{jk}^{-1} \end{array} \right.$ | $\underline{T}(r) = \underline{T}(r-1)\underline{F}(r)$ | $\underline{S}(r) = \underline{S}(r-1)\underline{G}(r)$ |

Note: $\{P_j\}$ is the distribution of population, $\{L_j\}$ is the distribution of land, and $\{d_{ij}\}$ the distribution of pairwise distances in the system. The matrices $\underline{F}(r)$ and $\underline{G}(r)$ are appropriately defined according to the logic presented in the text.

$\underline{G}(r)$ and their previous values $\underline{I}(r-1)$ and $\underline{S}(r-1)$ whereas the first two models use prior probability distributions consistent with the information-minimising strategy in equations (5.35) to (5.44).

The calibration of these three models was accomplished by solving equations (5.45) and (5.46) for $\mu_1$ and $\mu_2$ using the Newton-Raphson method. These models were particularly difficult to calibrate from an arbitrary starting point for $\mu_1$ and $\mu_2$. For example, in the case of the second model, a unique solution to equations (5.45) and (5.46) was obtained by starting with $\mu_1, \mu_2 = 0.01$, but with starting values of $\mu_1, \mu_2 = 0.2$, no solution was reached. This problem seems to be general to spatial interaction models with exponential or other functions such as Tanner's function (March, 1971) which involve more than one parameter. By successive substitution into equations (5.47) and (5.48) each interaction model can be expressed as a function of the initial prior and successive values of the parameters $\mu_1$ and $\mu_2$.

This implies that as in Tanner-type interaction models, the response surfaces of equations (5.45) and (5.46) are such that the optimum points are difficult to locate using iterative procedures, although global optima do exist. Thus, it is necessary to pick good starting values for any gradient method such as the Newton-Raphson, and this may be achieved using another method such as the Nelder-Mead simplex method (see Batty, 1976 for a detailed discussion). In fact, other versions of the model in which explicit attraction factors were used in equations (5.47) and (5.48) could not be calibrated using the Newton-Raphson method, and thus these attempts are not reported here.

The performance of these models depends very much upon the choice of

initial prior. The first and third models show an extremely poor performance in terms of the correlation between predicted and observed population, service employment, zonal activity rates and zonal population-serving ratios. In contrast, the second model which adopts an initial prior based on some knowledge of the activity distribution in the system performs quite reasonably although there is a tendency towards 'bogus' calibration which in this case involves the parameters cancelling one another out, and cancelling out the prior influence of distance.

Table 5.2 demonstrates the performance of these three models and it is of some interest to note that in terms of the zonal ratios, all three models perform badly. The third model is not worth illustrating in terms of explicit spatial predictions, but in Figure 5.4 the performance of the first model is illustrated using isometric smoothed surfaces of the observed and predicted population and service employment and their percentage deviation. This figure illustrates the difficulty of building a model which is just based on the dimensional properties of the space - land and distance, and on *a priori* grounds, it is not likely to give à good performance. Moreover, in this model, there is a tendency for the model to allocate too much activity to the periphery of the region and too little to the centre, as is illustrated by the percentage deviations.

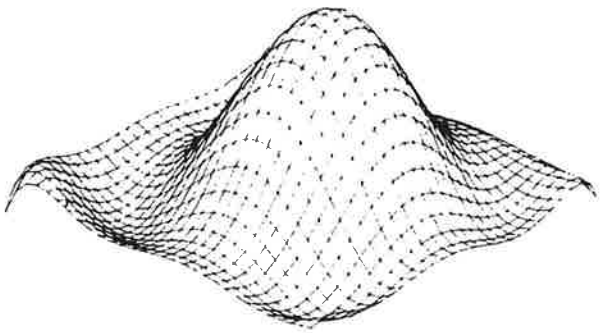
In Figure 5.5. these results are reversed. These surfaces show how good the performance of the second model is, and how this model tends to over-allocate activity at the centre of the region. As a final comment on these models, it does seem that the logic of updating an initial prior through the iterations of the model, is eminently reasonable, and is preferable on theoretical grounds to previous practice. The performance of a model of this kind too, is comparable to more traditional versions, and thus this line of research seems promising.

Table 5.2: Performance of the Activity Allocation Models.

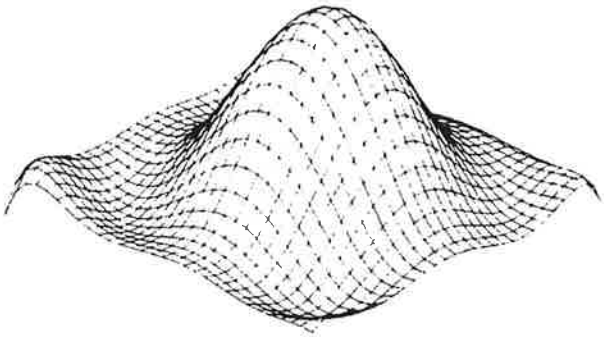
| Model | Parameters | | Correlations: $r^2$ | | | |
|---|---|---|---|---|---|---|
| | $\mu_1$ | $\mu_2$ | Population | Service employment | Zonal activity rates | Zonal population serving ratios |
| 1 | -0.0361 | 0.0773 | 0.4344 | 0.01537 | 0.5561 | 0.3212 |
| 2 | -0.0732 | 0.0743 | 0.9928 | 0.9116 | 0.6820 | 0.1102 |
| 3 | 0.0207 | 0.0329 | 0.2858 | 0.0763 | 0.2313 | 0.2953 |

Note: Mean work trip length = 9.8922 minutes; mean service trip length = 7.5318 ; parameters $\mu_1$ and $\mu_2$ started from 0.01 in the Newton-Raphson calibration method.

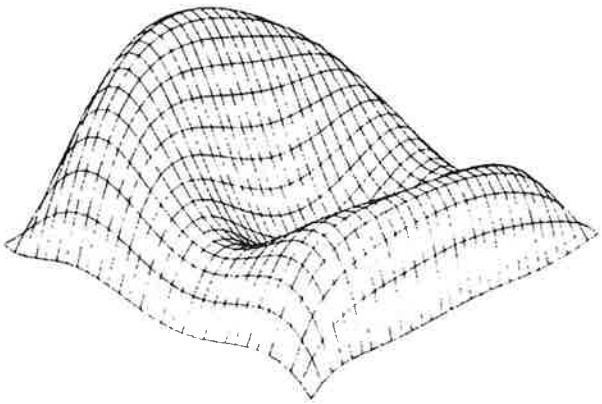
OBSERVED SERVICE EMPLOYMENT



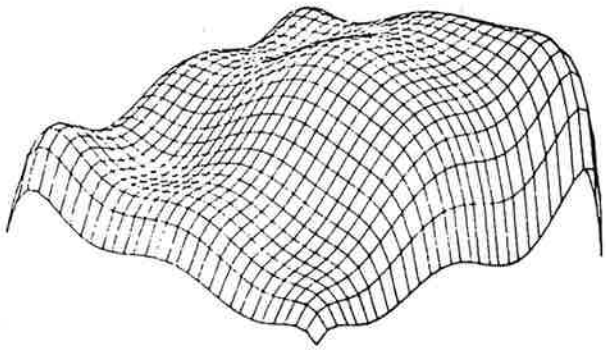
OBSERVED POPULATION



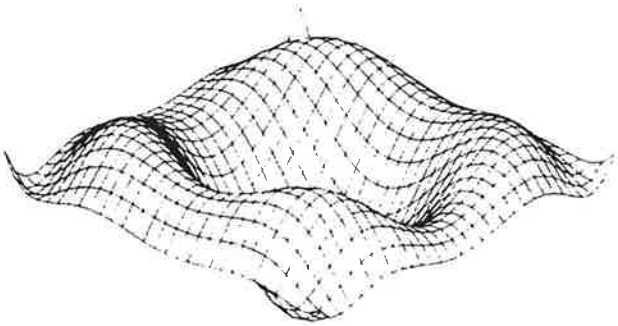
PREDICTED SERVICE EMPLOYMENT



PREDICTED POPULATION



PERCENTAGE DEVIATION IN SERVICE EMPLOYMENT



PERCENTAGE DEVIATION IN POPULATION

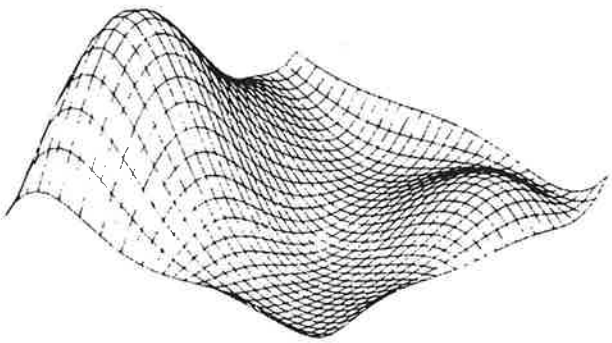
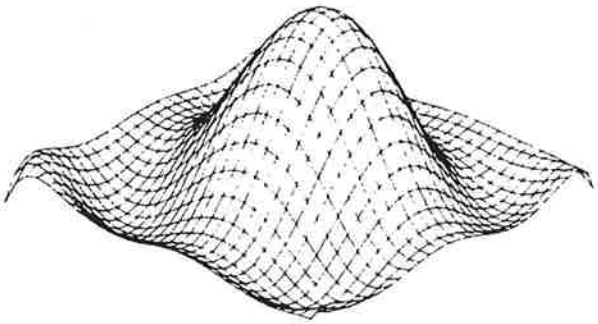
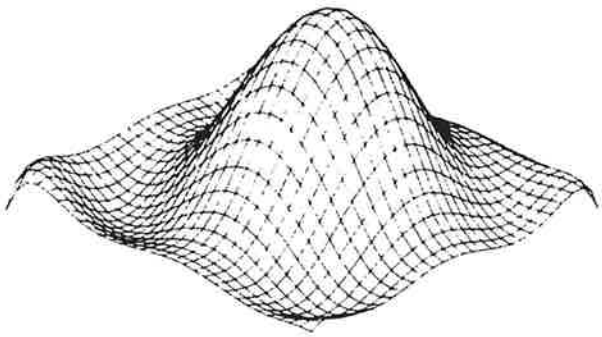


Figure 5.4: Spatial Predictions from the Model using a Prior based on Land and Distance.

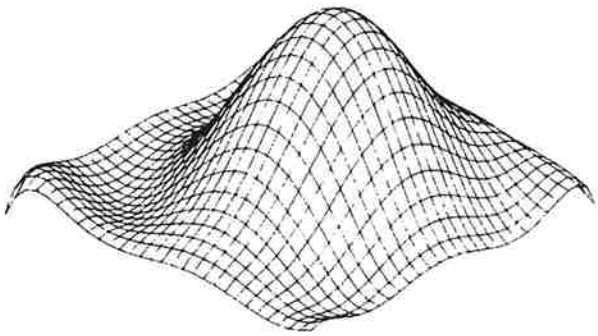
OBSERVED SERVICE EMPLOYMENT



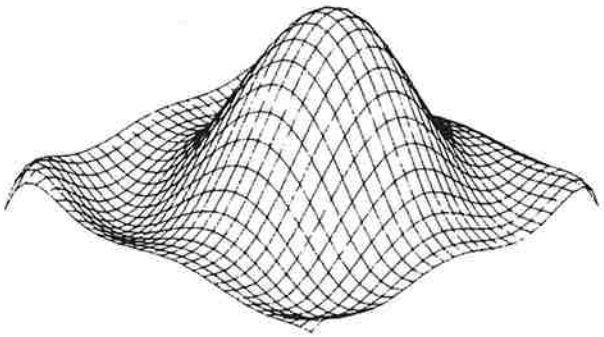
OBSERVED POPULATION



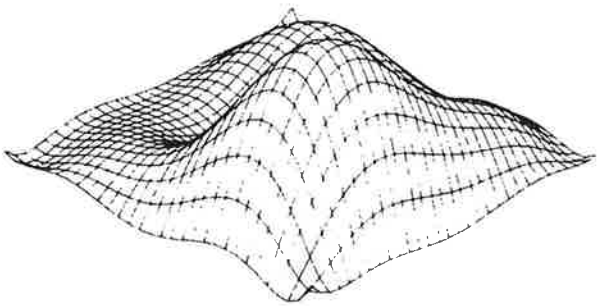
PREDICTED SERVICE EMPLOYMENT



PREDICTED POPULATION



DEVIATION PERCENTAGE IN SERVICE EMPLOYMENT



DEVIATION PERCENTAGE IN POPULATION

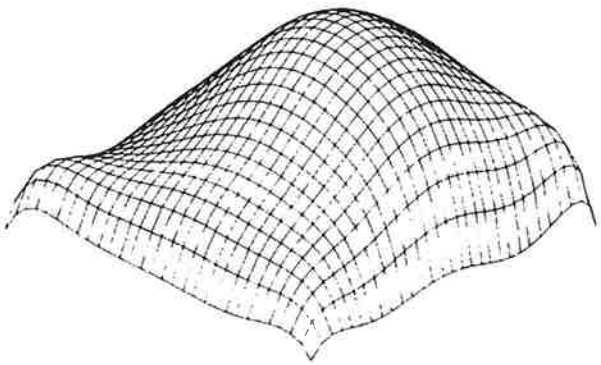


Figure 5.5: Spatial Predictions from the Model using a Prior based on Population and Distance.

CONCLUSIONS.

This chapter and the last have shown how the pseudo-dynamic model derived in Chapter 3 can be elaborated into a typology of different models, some of which are already known, some of which are quite new. In particular, the various elaborations of the Lowry model which integrate a dynamic economic base process with spatial interaction submodels, can be derived as special cases; the model due to Garin (1966) and Harris (1966), that due to Berechman (1976) and that made operational by Baxter and Williams (1975) are all models which can be reinterpreted in pseudo-dynamic terms. Yet the real power of the methodology comes in suggesting new model forms which are more relevant than existing practice from both a theoretical and practical standpoint. The notion of constraint processes being determined according to well-defined processes of redistribution, show how arbitrary existing constraint and calibration methods are, and the appeal of the pseudo-dynamic framework rests on the fact that new and more efficient procedures are immediately possible.

The central idea involving the use of the redistribution procedures for enabling constraints on location and interaction to be met, is based on the observation that all existing constraint procedures involve some element of trial-and-error (iteration). Thus the notion that the iterative solution of the model equations to satisfy the constraints be matched in some way to the iterative (dynamic) structure of the model itself, is inviting. Indeed, Chapter 4 which deals with the application of locational constraints in pseudo-dynamic models was concerned with establishing a correspondence between these processes, and there it was shown that constraints could always be met if solved according to the mover processes. Calibration of the interaction submodels, is also

effected iteratively in conventional models, and the suggestion that the calibration process be matched to the mover processes will be explored in the next two chapters. In a sense, both constraint and calibration can be interpreted as problems of optimal control and the analogy introduced here, is followed up in a more direct way in the next chapter.

Another aspect of the work reported here relates to the development of information-minimising in dynamic terms, and it appears that the constraint and calibration procedures might both be interpreted using this methodology. In this paper, the overall submodels applied to the Reading subregion, can be interpreted according to an information-minimising scheme, while the variation in these submodels through time reflects dynamic processes consistent with the overall structure, but not necessarily derivable using the formal methodology. The solution of the constraint equations relevant to any minimisation are usually iterative and a future task will be to find out whether this iterative solution can in turn be interpreted as an information-minimising process.

Finally, the models of this and the previous two chapters have a remarkable richness which can be exploited in many ways. The similarity to the educational models proposed by Stone (1970), for example, is striking, and suggests further interpretations which will be pursued elsewhere. As a way of closing the arguments of Chapters 3 to 5, Chapters 6 and 7 will examine the calibration problem in more detail, thus suggesting a new algorithm suitable for existing static models which can be interpreted in pseudo-dynamic terms. But by no means will this imply any degree of finality to these arguments for there are many directions which emerge from these ideas which will be taken up in later chapters.

# CHAPTER 6.

## COMPUTABLE MODEL FORMS BASED ON PSEUDO-DYNAMICS.

Urban systems are by necessity only observable at cross-sections in time, but any explanatory theory which seeks to unravel their structure must relate implicitly or explicitly to the processes which give rise to that structure. The degree to which these processes can be identified however, varies enormously and depends upon the existence of a suitable data base, upon the existence of intuitively acceptable hypotheses and upon the intrinsic nature of the observation itself. In fact, it appears easier to build static models which summarise the effects of such processes, rather than model the processes *per se*, but although static models tend to be the order of the day, it is still necessary to seek a greater understanding in terms of the dynamic processes at work. Rather than ignore such processes, it is possible to design urban models which contain both static and dynamic elements, or contain dynamic processes within more macro-static frameworks. In previous chapters, this idea was rationalised in the form of a pseudo-dynamic urban model: in Chapter 3, such a model was derived by aggregating a fully dynamic structure and in Chapters 4 and 5, the richness of the idea was demonstrated by a typology of such models, and their application.

Although pseudo-dynamic models represent approximations to dynamic processes viewed within a static framework, it is possible to exploit the idea in other ways. The fact that such dynamic processes exist within static models means that it is possible to use such processes to continually change, adapt or evolve the model to meet certain constraints. In short, pseudo-dynamic models can be controlled through their dynamic process in an analogous way to the engineer's use of feedback in the classical control theory of physical systems (Kendrick, 1976). Indeed, in the last chapter, such an idea was tentatively suggested for effecting a solution to the model which met prior locational constraints: as the model built up the system artificially through an economic-base type of process, constraints were checked at every stage, and a policy was initiated at each stage for resolving any constraint violation or for reaching a constraint in the subsequent stages.

In this chapter, this idea will be developed once again, but in relation to the problem of controlling or calibrating a small set of model parameters to meet some constraint on the patterns of spatial interaction predicted by the model. This problem is a fairly classical one in spatial model-building but in the past, indeed in the last chapter, it was treated in a static manner. This chapter attempts to demonstrate the idea that urban models with a pseudo-dynamic structure, can be more efficiently and sensibly calibrated if this structure is exploited.

The way in which this type of control can be accomplished depends upon the precise form of the pseudo-dynamic model, and as a starting point, it is necessary to briefly summarise the results of previous chapters. A pseudo-dynamic model was derived in Chapter 3 from a fully dynamic model characterised by two or more distinct time streams. By aggregating

at least one of these streams, and by leaving at least one in its fully dynamic form, the pseudo-dynamic model is derived. Associated with the model are different types of activity, namely new changes arising from exogenous inputs, movers who are relocating in response to changed locational structures, and stayers.

The new changes are generated from a constant and fixed input and the sequence generated from this input is geometrically convergent in the Leontief sense. However at each period of time in the sequence of generation, a new sequence involving movement in the original sequence can begin, and the last of these mover sequences starts in the period after the last new change has been generated. Thus the pseudo-dynamic model is mainly characterised by mover streams which initially depend upon the sequence of new change. In the last chapter, locational constraints were built into the process through the mover streams, whereas in this, the calibration of the system will be effected through both the new change and mover sequences.

This chapter does not however dwell entirely upon a process of evolutionary or adaptive calibration, for in the first sections, it is necessary to extend the results of the previous chapters by looking once again at the question of locational constraints. The pseudo-dynamic model of Chapters 3 and 4 is introduced first, slightly modified to deal with a variety of procedures of locational constraint, and then some typical procedures are outlined. Thus this chapter is fairly self-contained but readers should note the same caveat stated in the previous chapters: that the logic of the model depends upon ideas developed already, and thus previous chapters should be consulted. The problem of treating locational constraints is one of developing an efficient but consistent

computational form, and although the complete pseudo-dynamic model is consistent, it is computationally excessive. Therefore, a simpler form of model is developed here in which the mover and new change activity streams are collapsed, and the recursive form for this model provides the structure developed here.

The argument then changes in pitch, and the focus in the rest of this chapter and the next is on calibration. A non-matrix presentation of the model and its interaction submodels is given and the calibration problem is explored in conceptual terms. An algorithm in which the system parameters are adjusted to meet the trip lengths required is sketched in the next chapter and certain aspects are elaborated: the evaluation of required trip lengths, the establishment of feasible bounds on the process, and the directions of search needed are described. Finally, the algorithm is tested on a model of the Peterborough urban region, and certain conclusions as to its efficiency are drawn. In developing these ideas, several themes for future research have been evolved, and by way of conclusion, a programme for future work is suggested.

AN OUTLINE OF THE PSEUDO-DYNAMIC MODEL.

The starting point in this paper is a statement of the pseudo-dynamic model given previously in equations (4.1) to (4.7). The notation is the same as in previous chapters but it will be reintroduced here so that readers focussing just on the calibration of this model will find this chapter together with the next self-contained. Three major time periods characterise the model: the input of activity in period [t-T:t-T-1] is associated with new change only, the generation of further activity from the input and the movement of activity already generated occurs in

the period [t:t-T] and only the movement of existing activity occurs in the period [t+T+1:t]. In fact, each sequence of new change and movers has a life of T time units, that is, if the input is associated with the first unit of new change in [t-T:t-T-1], the last unit of new change generaged is T+1 time periods later in [t:t-1]. Note that the period indexed by the time script $\tau$ refers to the unit of time in the interval $\tau-(\tau-1)=1$. The first major period relates to only new change, the second to a build up of the system through movers and new change, and the third to a period of decline in momentum as mover streams terminate to a stable situation at t+T+1.

The model will be developed for population and employment first at a macro-level, and then at a more micro-level in terms of its submodels governing the sequence of generation and distribution. Population and employment are described by 1 X N row vectors $\underline{p}(\tau)$ and $\underline{e}(\tau)$ respectively and the convention that bold lower case letters indicate 1 x N vectors and bold upper case letters N x N matrices is adhered to. The initial input of employment which drives the model is defined as $\Delta^*\underline{s}(0)$ which is the first of a series of increments of new employment change, $\Delta^*\underline{s}(\tau)$, and new population change, $\Delta^*\underline{p}(\tau)$.

Then in the first major time period [t-T:t-T-1],

$$\underline{p}(t-T) = \Delta^*\underline{p}(t-T), \quad \text{and} \tag{6.1}$$

$$\underline{e}(t-T) = \Delta^*\underline{s}(0) + \Delta^*\underline{s}(t-T). \tag{6.2}$$

In the second major period [t:t-T], the state equations of the model are composed of movers and stayers as well as new change. The appropriate

variable is superscripted by m or s according to whether it reflects movers or stayers and the variable is postscripted according to historical time r, and the time when the new change activity from which the movers originate, was first generated, time w. Then

$$\underline{p}(r) = \Delta^*\underline{p}(r) + \sum_{w=t-T}^{r-1} \underline{p}^m(r,w)\underline{\Psi}(r,r-u) + \sum_{w=t-T}^{r-1} \underline{p}^s(r,w), \qquad (6.3)$$

where the index r-u denotes the time when the mover sequence originated, and u=w-(t-T) = w-t+T. For employment

$$\underline{e}(r) = \Delta^*\underline{s}(0) + \Delta^*\underline{s}(r) + \sum_{w=t-T}^{r-1} \underline{s}^m(r,w)\underline{\Psi}(r,r-u) + \sum_{w=t-T}^{r-1} \underline{s}^s(r,w). \qquad (6.4)$$

The matrix $\underline{\Psi}(r,r-u)$ is a switch function which has been introduced here to indicate whether the appropriate term in the mover sequence is relevant or not. In essence, it determines which part of the mover sequence is operative and it relates to the various constraint procedures outlined below. When the switch is on, $\underline{\Psi}(r,r-u) = \underline{I}$ and when it is off, $\underline{\Psi}(r,r-u)=\underline{0}$. Figure 6.1 shows the typical streams associated with this model, and whether or not any part of any mover stream is being used, is controlled by the switch.

The third major period from t+1 to t+T+1 does not contain any new change for this sequence which starts at t-T, ends at t, and thus only movers and stayers feature in the state equations. Also as the mover streams are beginning to terminate in this period, the stayers can be divided into two groups: those who are still associated with activity still moving, and those whose activity base is stable. Then the population equation is

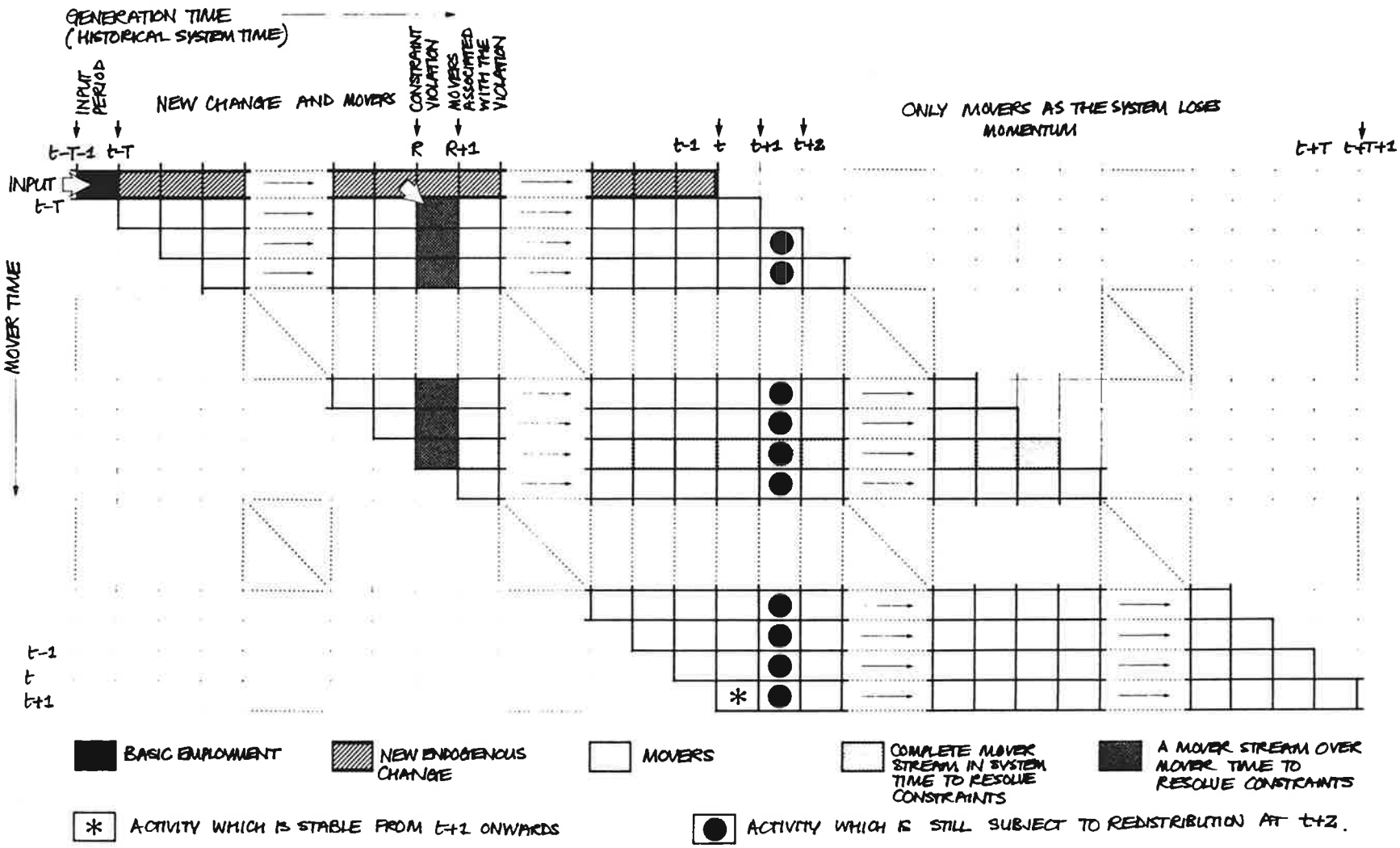



Figure 6.1: Time Streams Characterising the Pseudo-Dynamic Model.

$$\underline{p}(r) = \sum_{w=r-T-1}^{t} \underline{p}^{m}(r,w)\underline{\Psi}(r,r-u) + \sum_{w=r-T-1}^{t} \underline{p}^{s}(r,w) + \sum_{w=t-T}^{r-T-2} \underline{p}^{s}(w+T+1,w), \tag{6.5}$$

and the employment equation

$$\underline{e}(r) = \Delta^{*}\underline{s}(0) + \sum_{w=r-T-1}^{t} \underline{s}^{m}(r,w)\underline{\Psi}(r,r-u) + \sum_{w=r-T-1}^{t} \underline{s}^{s}(r,w) + \sum_{w=t-T}^{r-T-2} \underline{s}^{s}(w+T+1,w). \tag{6.6}$$

Note that the last terms in equations (6.5) and (6.6) are out of range when r=t+1 and are thus undefined. This is clear when the column associated with t+1 in Figure 6.1 is compared to the t+2 column where the unit of change associated with the first unit of new change generated is now stable. This is because the last sequence of movers begins at t+1 and the first unit of this sequence moving at t+1, is stable thereafter.

A RECURSIVE FORM FOR THE MODEL.

It is now necessary to develop models of the various components in equations (6.1) to (6.6) which relate the state equations in recursive fashion. The relationship between the two equations through time is sequential, starting with an employment input generating population, generating more employment, more population and so on. The sequence of generation involves both the derivation of one activity from another in a geometrically convergent manner, and their allocation in space using spatial interaction models. New change is the easiest to handle: first for new population,

$$\Delta^{*}\underline{p}(r) = \Delta^{*}\underline{s}(r-1)\underline{A}(r), \tag{6.7}$$

where $\underline{A}(r)$ is an N x N distribution matrix which maps the set of employment locations into population locations as well as scaling employment to population. As in previous chapters, it is assumed that $\underline{A}(r)$ is separable in the following sense

$$\underline{A}(r) = \underline{T}(r)\underline{\Lambda} ,$$

where $\underline{T}(r)$ is an N x N row stochastic probability matrix linking work to home and $\underline{\Lambda}$ is a scalar diagonal matrix of inverse activity rates $\lambda$.

In a similar way,

$$\Delta^*\underline{s}(r) = \Delta^*\underline{p}(r)\underline{B}(r), \tag{6.8}$$

where $\underline{B}(r)$ is separable into

$$\underline{B}(r) = \underline{\Gamma}\ \underline{S}(r).$$

$\underline{\Gamma}$ is a scalar diagonal matrix of population-service demand ratios $\gamma$ and $\underline{S}(r)$ is a row stochastic probability matrix associated with the spatial pattern of demand for services. Substituting for $\Delta^*\underline{p}(r)$ in equation (6.8) from (6.7) and expressing in separable form leads to

$$\Delta^*\underline{s}(r) = \Delta^*\underline{s}(r-1)\underline{T}(r)\underline{\Lambda}\ \underline{\Gamma}\ \underline{S}(r), \tag{6.9}$$

which is the central recurrence relation of new change in the model. Then from the initial input $\Delta^*\underline{s}(0)$

$$\Delta^*\underline{p}(r) = \Delta^*\underline{s}(0)[\underline{\Lambda}\ \underline{\Gamma}]^{r-t+T}\underline{\Lambda} \prod_{\tau=t-T}^{r-1} \underline{T}(\tau)\underline{S}(\tau)\underline{T}(r), \tag{6.10}$$

$$\Delta^*\underline{s}(r) = \Delta^*\underline{s}(0)[\underline{\Lambda}\ \underline{\Gamma}]^{r-t+T+1} \prod_{\tau=t-T}^{r} \underline{T}(\tau)\underline{S}(\tau). \tag{6.11}$$

From equations (6.10) and (6.11), it is clear that the convergence depends upon the scalar diagonal matrix $\underline{\Lambda}\ \underline{\Gamma}$ where it is necessary for $\underline{0}<\underline{\Lambda}\ \underline{\Gamma}<\underline{I}$ for non-trivial urban systems.

The sequence of generation for any mover stream is similar to that for new change apart from the fact that a proportion $\underline{\alpha}(r,r-u)$ of the new change is moved, $\underline{0} \leq \underline{\alpha}(r,r-u) \leq \underline{I}$, and the matrix $\underline{\alpha}(r,r-u)$ is now assumed to be scalar diagonal. Then in a manner similar to above, population and employment (service) movers are generated by

$$\underline{p}^m(r,w) = \underline{s}^m(r-1,w-1)\underline{\alpha}^{-1}(r-1,r-u)\underline{\alpha}(r,r-u)\underline{\tilde{A}}(r), \quad \text{and} \tag{6.12}$$

$$\underline{s}^m(r,w) = \underline{p}^m(r,w)\underline{\tilde{B}}(r) \tag{6.13}$$

The matrices $\underline{\tilde{A}}(r)$ and $\underline{\tilde{B}}(r)$ reflect changed patterns of distribution which is one reason for initiating the mover sequence in the first place, and these matrices are separable in the same sense as those pertaining to the new change. That is

$$\underline{\tilde{A}}(r) = \underline{\tilde{T}}(r)\underline{\Lambda}, \quad \text{and}$$

$$\underline{\tilde{B}}(r) = \underline{\Gamma}\,\underline{\tilde{S}}(r).$$

The basic recurrence relation for movers can now be derived as

$$\underline{s}^m(r,w) = \underline{s}^m(r-1,w-1)\underline{\alpha}^{-1}(r-1,r-u)\underline{\alpha}(r,r-u)\underline{\tilde{T}}(r)\underline{\Lambda}\,\underline{\Gamma}\,\underline{\tilde{S}}(r), \tag{6.14}$$

and note that the boundary conditions for each mover stream are stated as $\underline{s}^m(r-1,r-1) = \underline{0}$, $\underline{p}^m(r-1,r-1) = \underline{0}$. Then in terms of the original input $\Delta^*\underline{s}(0)$, recursion on equations (6.12) and (6.13) leads to

$$\underline{p}^m(r,w) = \Delta^*\underline{s}(0)\underline{\alpha}(r,r-u)[\underline{\Lambda}\,\underline{\Gamma}]^u\underline{\Lambda}\prod_{\tau=r-u}^{r-1}\underline{\tilde{T}}(\tau)\underline{\tilde{S}}(\tau)\underline{\tilde{T}}(r), \tag{6.15}$$

$$\underline{s}^m(r,w) = \Delta^*\underline{s}(0)\underline{\alpha}(r,r-u)[\underline{\Lambda}\,\underline{\Gamma}]^{u+1}\prod_{\tau=r-u}^{r}\underline{\tilde{T}}(\tau)\underline{\tilde{S}}(\tau)\ . \tag{6.16}$$

It is also clear that each mover sequence converges for non-trivial problems due to $\underline{0} \leq \underline{\alpha}\,(r,r-u)[\underline{\Lambda}\,\underline{\Gamma}] \leq [\underline{\Lambda}\,\underline{\Gamma}] \leq \underline{I}$, and the rate of convergence is as fast, if not faster than the original sequence of new change.

The final sequence relates to the stayers and as these are basically a function of the movers, it is only necessary to state the usual recurrence relations developed in Chapter 3. Then

$$\underline{p}^s(r,w) = [\underline{p}^m(r-1,w)\underline{\Psi}(r-1,r-u-1)+\underline{p}^s(r-1,w)][\underline{I}-\underline{\alpha}(r,r-u)], \qquad (6.17)$$

$$\underline{s}^s(r,w) = [\underline{s}^m(r-1,w)\underline{\Psi}(r-1,r-u-1)+\underline{s}^s(r-1,w)][\underline{I}-\underline{\alpha}(r,r-u)], \qquad (6.18)$$

where $\underline{s}^s(r-1,r-1) = \Delta^*\underline{s}(r-1)$ and $\underline{p}^m(r-1,r-1) = \Delta^*\underline{p}(r-1)$ are the boundary conditions. It is now possible to substitute the specific submodel forms developed in equations (6.7) to (6.18) into equations (6.3) to (6.6) to derive the specific state of the system as a function of its previous state. The equations are fairly lengthy and as the population state equation is a subset of the employment equation, only the employment equation need be stated, This was a convention adopted in previous chapters and it will be used here when there is no ambiguity.

For the period [t+1:t-T]

$$\underline{e}(r) = \Delta^*\underline{s}(0)+\{\Delta^*\underline{s}(r-1)\underline{T}(r)\underline{S}(r)+[\Delta^*\underline{s}(0)\underline{\alpha}(r,r)\underline{\Psi}(r,r)$$

$$+ \sum_{w=t-T+1}^{r-1} \underline{s}^m(r-1,w-1)\underline{\alpha}^{-1}(r-1,r-u)\underline{\alpha}(r,r-u)\underline{\Psi}(r,r-u)]\tilde{\underline{T}}(r)\tilde{\underline{S}}(r)\}\underline{\Lambda}\ \underline{\Gamma}$$

$$+ \sum_{w=t-T}^{r-1} [\underline{s}^m(r-1,w)\underline{\Psi}(r-1,r-u-1)+\underline{s}^s(r-1,w)][\underline{I}-\underline{\alpha}(r,r-u)]. \qquad (6.19)$$

Equation (6.19) refers to the period up to t+1 which is one time period longer than equation (6.4). This is due to the fact that only after t+1 does the term involving the initiation of movers from the initial input cease to exist. Furthermore, the recurrence relation on movers

used in the equation after t+1 does not hold for t+1, thus equation (6.19) must be used instead. Note that when r=t+1 in equation (6.19), $\Delta^*\underline{s}(t+1) = \underline{0}$. In fact, the time period [t+1:t] is somewhat anomalous in that either the second major period or third major period equation will usually hold as noted in the previous paper.

Then for the period [t+T+1:t+1]

$$\begin{aligned}\underline{e}(r) &= \Delta^*\underline{s}(0) \\ &+ \sum_{w=t-T+1}^{t} \underline{s}^m(r-1,w-1)\underline{\alpha}^{-1}(r-1,r-u)\underline{\alpha}(r,r-u)\underline{\Psi}(r,r-u)\underline{\tilde{T}}(r)\underline{\Lambda}\ \underline{\Gamma}\ \underline{\tilde{S}}(r) \\ &+ \sum_{w=r-T-1}^{t} [\underline{s}^m(r-1,w)\Psi(r-1,r-u-1)+\underline{s}^s(r-1,w)][\underline{I}-\underline{\alpha}(r,r-u)] \\ &+ \sum_{w=t-T}^{r-T-2} \underline{s}^s(w+T+1,w). \qquad (6.20)\end{aligned}$$

This completes the statement of the pseudo-dynamic model. Before a form suitable for adaptive calibration is derived, it is worthwhile exploring further ways in which locational constraints might be incorporated in this form, for this is of crucial importance in deriving a structure suitable for efficient calibration.

PROCEDURES FOR INCORPORATING LOCATIONAL CONSTRAINTS.

By combining the switch function $\underline{\Psi}(r,r-u)$ and the mover ratio matrix $\underline{\alpha}(r,r-u)$ in various ways, many different types of constraint procedure can be developed for the pseudo-dynamic model. The easiest set of models to generate are those within the typology developed in Chapter 4: in the case of $\underline{\alpha}(r,r-u) = \underline{0}$ models, the switch is redundant but it might

be set equal to $\underline{0}$ for completeness; in the case of $\underline{\alpha}(r,r-u) = \underline{\alpha}$ or $\underline{\alpha}(r,r-u) = \underline{I}$ models, the switch is always on at $\underline{\Psi}(r,r-u) = \underline{I}$. In terms of constraint procedures, the $\underline{\alpha}(r,r-u)$ matrix controls the amount of activity which is moved in an effort to overcome a constraint violation or to meet a constraint whereas the switch function $\underline{\Psi}(r,r-u)$ controls the elements of the mover sequences which are relevant to the relocation of existing activity. Although mover sequences are activated in the event of some constraint violation, the form of which must be determined in advance, the type of constraint procedure must also be specified *a priori*.

In the previous two chapters, rather strict procedures were specified in terms of the switch function. Complete sequence redistribution specified by $\underline{\Psi}(r,r-u) = \underline{I}$, $r>r-u$ and partial sequence redistribution where $\underline{\Psi}(r,r-u) = \underline{0}$ for some part of the generation sequence given by $w' \leq r$ and $\underline{\Psi}(r,r-u) = \underline{I}$, $w'>r$, were both described; but an *ad hoc* possibility exists which has complete flexibility where the particular element of the sequence is switched on when necessary to solve a constraint violation. Yet even in this case, some idea of the form of the procedure must be specified before the simulation begins.

To operate the complete and partial sequence procedures of the previous chapter, assume that the first constraint violations occur at time R, and thus initiate a mover sequence beginning in every period $r \geq R+1$. Then prior to R+1, the switch is off: for complete sequences $\underline{\Psi}(r,r-u)=\underline{0}$, $r \leq R$, $t-T \leq w \leq t$, and for partial sequences $\underline{\Psi}(r,r-u)=\underline{0}$, $r \leq R$, $w<w'$. After the constraint violation, the opposite conditions hold, but note that to operate the model as stated in the previous section, it is necessary to assume $\underline{\alpha}(r,r-u) = \underline{I}$, $r \leq R$, $w<w'$ for the partial sequence so that the

mover recurrence relations can be applied. In the *ad hoc* situation however, it is necessary to assume that the switch function dominates the process: that is, the mover ratio is always set as $\underline{\alpha}(r,r-u) = \underline{I}$ and $\underline{\Psi}(r,r-u) = \underline{0}$ whenever the element in the sequence is inoperative but that $\underline{\alpha}(r,r-u)$ is set to its appropriate value and $\underline{\Psi}(r,r-u) = \underline{I}$ when the element is operative. This requirement to set $\underline{\alpha}(r,r-u) = \underline{I}$ is necessitated by the form of the mover recurrence relations given in equations (6.12) to (6.14).

The actual procedure in which constraint violations are resolved through mover sequences is also flexible. The problem can be treated as a constrained matrix problem in the manner developed for transport models (Macgill, 1975) or it can be treated as one of redistribution of the surplus in a manner akin to that originally used by Lowry (1964) and developed by Echenique, Crowther and Lindsay (1969). In Chapters 8 and 9, the role of constrained matrix methods will be examined in the context of pseudo-dynamic models but here the more arbitrary surplus redistributing procedure will be developed. The idea behind this latter procedure was outlined in Chapter 5 and it will be introduced again here in a rather different way, first for complete sequence mover streams, and then for the more *ad hoc* procedure presented in the next section.

The central core of the constraint algorithm is based on the concept of assessing the greatest constraint violation based on population and employment, and using this to determine the mover ratio. Then given constraint vectors $\underline{c}^p(r)$ on population and $\underline{c}^e(r)$ on employment, defining the set of constrained population zones as $Z_p$ and employment zones as $Z_e$, the procedure is operated as follows. Assume a constraint is violated at time R: then if

$$P_j(R) \geq C_j^p(R) \ , \quad j \text{ is assigned to } Z_p, \tag{6.21}$$

and the surplus $\Delta_j^p(R)$ is computed as

$$\Delta_j^p(R) = P_j(R) - C_j^p(R), \ j\varepsilon Z_p. \tag{6.22}$$

The ratio to be redistributed, $\rho(R+1,R+1-u)$, is taken as a proportion $\phi(R+1-u)$ of the surplus

$$\rho(R+1,R+1-u) = \phi(R+1-u) \sum_{j\varepsilon Z_p} \Delta_j^p(R), \tag{6.23}$$

and this proportion depends upon the type of procedure used which is outlined below.

At this point, the trip probability $\tilde{t}_{ij}(R+1)$ is renormalised

$$\tilde{t}_{ij}(R+1) = 0, \quad j\varepsilon Z_p, \tag{6.24}$$

and the algorithm turns to deal with employment violations. If

$$E_k(R) \geq C_k^e(R), \quad k \text{ is assigned to } Z_e, \tag{6.25}$$

and the surplus $\Delta_k^e(R)$ is computed as

$$\Delta_k^e(R) = E_k(R) - C_k^e(R), \ k\varepsilon Z_e \ . \tag{6.26}$$

Another ratio $\sigma(R+1,R+1-u)$ is formed as a proportion $\theta(R+1-u)$ of the surplus, and this is given by

$$\sigma(R+1,R+1-u) = \theta(R+1-u) \sum_{k\varepsilon Z_e} \Delta_k^e(R), \tag{6.27}$$

and the trip probability $\tilde{s}_{jk}(R+1)$ is renormalised as

$$\tilde{s}_{jk}(R+1) = 0, \qquad k\varepsilon Z_e \ . \tag{6.28}$$

At this point, the mover ratio matrix $\underline{\alpha}(R+1,r+1-u)$ and the switch function $\underline{\Psi}(R+1,R+1-u)$ must be set for R+1. Then

$$\alpha_{ij}(R+1,R+1-u) \begin{cases} = \max[\sigma(R+1,R+1-u),\rho(R+1,R+1-u)], \ i=j, \\ \\ = 0, \ \ i \neq j, \end{cases} \tag{6.29}$$

and

$$\Psi_{ij}(R+1,R+1-u) \begin{cases} = 1, \ \ i=j, \\ \\ = 0, \ \ i\neq j. \end{cases} \tag{6.30}$$

Note that the mover ratio matrix is scalar diagonal and that the parameters $\phi(R+1-u)$ and $\rho(R+1-u)$ are determined by the particular type of constraint procedure.

The method of Chapter 5 in which a complete mover sequence was initiated due to a constraint violation can be treated in these terms. In that method, the various ratios were independent of generation time w and associated with a mover stream was a constant ratio fixed at time R. Then

$$\theta(R+1) = (1-\gamma\lambda)/\sum_i \Delta^* S_i(0), \qquad \text{and}$$

$$\phi(R+1) = \lambda^{-1}\theta(R+1).$$

The mover ratios and switch functions are set at R as

$$\alpha_{ij}(r,R+1) \begin{cases} = \max\,[\sigma(R+1),\ \rho(R+1)], & i=j, \\ = 0, & i\neq j, \end{cases}$$

$$\Psi_{ij}(r,R+1) \begin{cases} = 1, & i=j, \\ = 0, & i\neq j, \end{cases}$$

and the range of r is $R+1 \leq r \leq R+T+1$. The mover stream thus depends on a value of $\underline{\alpha}(r,R+1)$ fixed at R and the surplus is redistributed according to the whole of the activity associated with such a sequence. Diagramatically, this implies that the stream which is represented by the light stipple in Figure 6.1 is utilised to redistribute this surplus.

Another method is based on fixing the proportions in terms of the activity generated so far, or in terms of the activity still being moved. In essence, this method involves redistributing the surplus in terms of all the activity generated so far, and this is achieved immediately at R+1, rather than from R+1 to R+T+1 as in the previous method. Then

$$\theta(R+1,R+1-u) = \sum_k \Delta^* S_k(w) / [\sum_{\tau=t-T}^{R} \sum_k \Delta^* S_k(\tau)]^2, \qquad \text{and}$$

$$\phi(R+1,R+1-u) = \sum_j \Delta^* P_j(w) / [\sum_{\tau=t-T}^{R} \sum_j \Delta^* P_j(\tau)]^2.$$

These equations refer to the period [t+1:t-1] and after t+1, the range of summation for $\tau$ in these equations is $R-T-1 \leq \tau \leq t$, due to the fact that the last sequence of movers begins at t+1. In Figure 6.1, the activity allocation of this surplus is over all the mover streams associated with time R+1, and this is shown by the dark stipple in contrast to the above procedure.

One immediate and obvious problem of both these and other constraint procedures outlined in this section is computational feasibility. The model of the Reading region outlined in Chapter 5 was feasible in that storage requirements were kept down by assuming $\underline{A}(\tau) = \underline{\tilde{A}}(\tau)$, $\underline{B}(\tau) = \underline{\tilde{B}}(\tau)$ and $\underline{\alpha}(r,r-u)$ dependent only on r. But here there is another dimension of complexity to be added and this pertains to calibration. Additional storage will be necessary for the algorithm developed below, and thus a more efficient, more parsimonious form of constraint procedure is required. This is developed in the following section.

A SIMPLIFIED FORM FOR THE URBAN MODEL.

The simplest form of pseudo-dynamic model in which movers are represented, is the model in which there is only one mover stream. In this case, it might be assumed that r-u = t-T+1 which implies that r and w are equivalent. Each component in the mover stream is associated with a ratio $\underline{\alpha}(r,t-T+1)$ and the switch $\underline{\Psi}(r,t-T+1)$ is always on. That is

$$\underline{\Psi}(r,r-u) \begin{cases} = \underline{I}, & r-u = t-T+1, \\ = \underline{0} & r-u \neq t-T+1. \end{cases}$$

The model is worth developing explicitly if only to show how even this form has disadvantages which must be resolved by the model to be developed here. Then for the employment equations, in the period [t:t-T]

$$\underline{e}(r) = \Delta^*\underline{s}(0) + \Delta^*\underline{s}(r) + \underline{s}^m(r,r-1) + \sum_{w=t-T}^{r-1} \underline{s}^s(r,w), \qquad (6.31)$$

which is derived from equation (6.4) given information about the switch function. $\Delta^*\underline{s}(r)$ is as given in equation (6.11) but the movers and stayers have a particularly simple form. Then

$$\underline{s}^m(r,r-1) = \Delta^*\underline{s}(r-2)\underline{\alpha}(r,t-T+1)\underline{\tilde{A}}(r)\underline{\tilde{B}}(r), \tag{6.32}$$

$$\underline{s}^s(r,r-1) = \Delta^*\underline{s}(r-1)[\underline{I}-\underline{\alpha}(r,t-T+1)] , \tag{6.33}$$

and from the recurrence relation on movers and stayers in equations (6.14) and (6.18) respectively,

$$\begin{aligned}\underline{s}^s(r,w) &= \Delta^*\underline{s}(w-1)\underline{\alpha}(w+1,t-T+1)\underline{\tilde{A}}(w+1)\underline{\tilde{B}}(w+1) \quad + \\ &\quad \Delta^*\underline{s}(w)[\underline{I}-\underline{\alpha}(w+1,t-T+1)], \quad w<r-1.\end{aligned} \tag{6.34}$$

Using equations (6.31) to (6.34), the employment equation in (6.31) can be written

$$\begin{aligned}\underline{e}(r) &= \Delta^*\underline{s}(0) + \Delta^*\underline{s}(r) \\ &+ \sum_{\tau=t-T}^{r-2} \Delta^*\underline{s}(\tau)\underline{\alpha}(\tau+2,t-T+1)\underline{\tilde{A}}(\tau+2)\underline{\tilde{B}}(\tau+2) \\ &+ \Delta^*\underline{s}(0)\underline{\alpha}(t-T+1,t-T+1)\underline{\tilde{A}}(t-T+1)\underline{\tilde{B}}(t-T+1) \\ &+ \sum_{\tau=t-T}^{r-1} \Delta^*\underline{s}(\tau)[\underline{I}-\underline{\alpha}(\tau+1,t-T+1)] .\end{aligned} \tag{6.35}$$

An equation for the third major period is not necessary. As there is only a first order lag between new change and movers, the model terminates at t+1, and equation (6.35) applies for t+1 with $\Delta^*\underline{s}(t+1) = \underline{0}$. Figure 6.2(a) shows this sequence and the direct relationships between movers and new change. From equation (6.35), it is clear that the model is based directly on new change and although this is clearly the case with models based on many mover streams, the one mover stream model is sufficiently simple to represent in terms of only new change. The implication of equation (6.35) is that after each constraint violation, the appropriate portion of the single mover stream is utilised to resolve this violation. This is the essence of Figure 6.2(a).

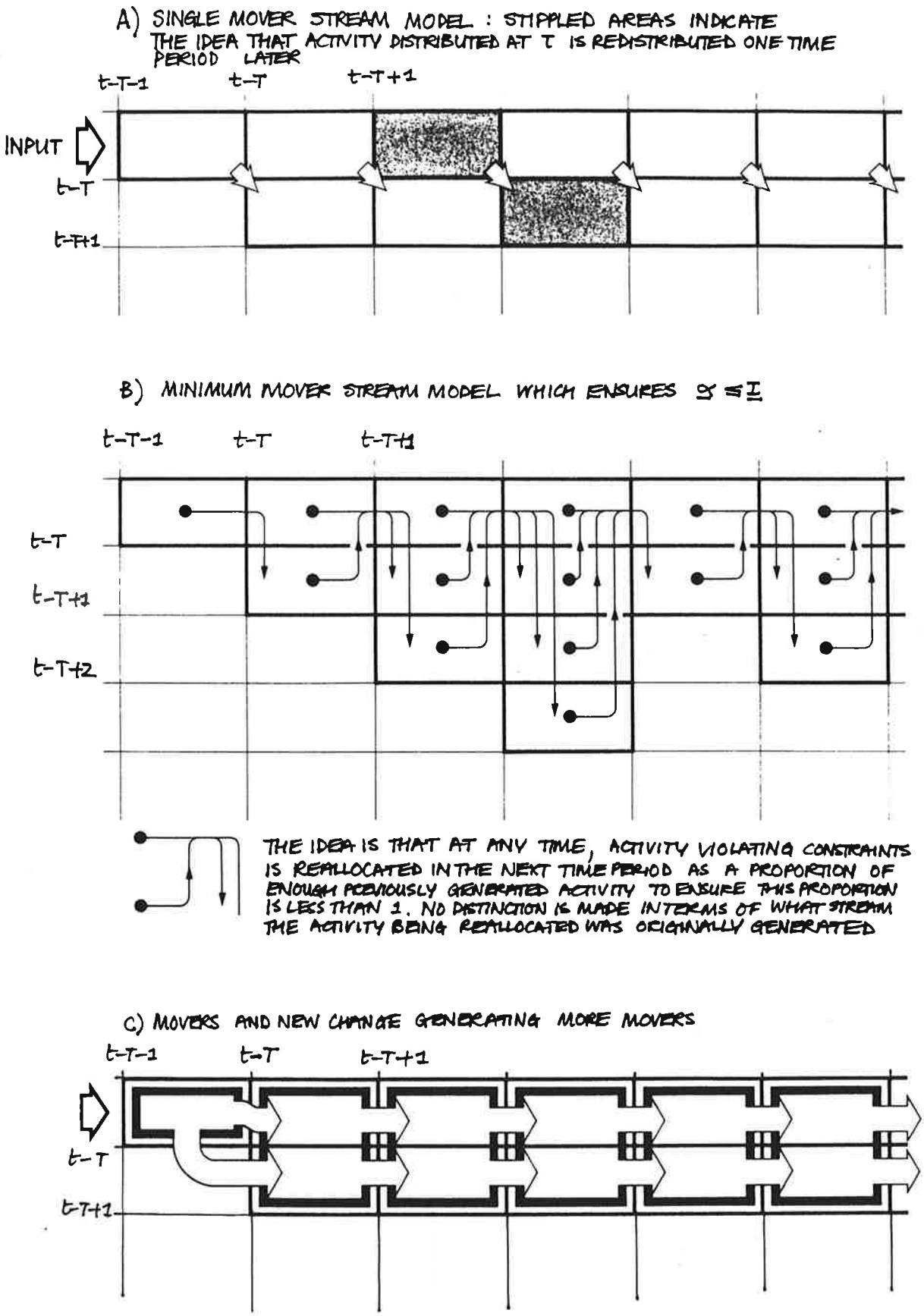



Figure 6.2: Limited Mover Stream Models Designed to Resolve the Locational Constraint Problem.

The problem with the model as stated relates to whether or not the constraint violation can be resolved by moving only one element of the mover sequence at a time. Noting that R=w and that R+1-u = t-T+1, the proportions $\theta(R+1-u)$ and $\phi(R+1-u)$ are given by

$$\theta(R+1-u) = 1/\sum_k \Delta^* S_k(R), \quad \phi(R+1-u) = 1/\sum_j \Delta^* P_j(R).$$

Thus the amount moved is taken as a proportion of the previous new activity generated and thus it is clear from equation (6.32) that all the surplus is redistributed in the next time period. Clearly there is no guarantee that this redistribution will meet all the constraints: it will certainly meet those which have already been violated but it may not meet those which have not yet been infringed.

In other words, it is possible for any time period after the time when the first constraints were violated at R for

$$\sum_j \Delta_j^p(r) > \sum_j \Delta^* P_j(r), \quad \text{and/or,} \quad \sum_k \Delta_k^e(r) > \sum_k \Delta^* S_k(r), \quad r \geq R+1$$

Such a situation would involve the mover ratio $\underline{\alpha}(r+1,t-T+1)$ exceeding the identity matrix, and although this would be perfectly permissible, it would demonstrate an inconsistency in the single mover stream method. This inconsistency relates to the fact that the constraint violation is not being resolved immediately in the following time period but is being perpetuated through time. Of course, by the last iteration of the simulation, all the constraints would be satisfied.

If for consistency, it is required to keep the mover ratio matrix less than $\underline{I}$, then the single mover stream idea must be abandoned. In fact,

it may be necessary to expand the mover streams in a somewhat *ad hoc* way, so that enough activity is included in the proportions $\theta(R+1-u)$ and $\phi(R+1-u)$ for $\underline{\alpha}(R+1,t-T+1) \leq \underline{I}$. Figure 6.2(b) illustrates the way in which the mover streams begin to spread as the need to keep $\underline{\alpha}(r,t-T+1)$ less than $\underline{I}$ is met, and it also indicates the relative difficulty with any schemes for resolving constraints which are based on *ad hoc* elements from the mover sequences. Moreover, Figure 6.2(b) shows the importance of 'packing' the mover sequences as close together as possible so that computational storage and effort is conserved. Indeed, it is interesting to think of the problem of designing efficient constraint procedures in models of this type as 'packing problems'. The advantage of this discussion is based on the insights into the problem of constraints in single mover stream models: if it is necessary to retain the single mover stream idea, if iterationwithin each time period is not to be allowed and if $\underline{\alpha}(r,t-T+1) \leq \underline{I}$, then it is necessary to design a different model from the one described in equations (6.31) to (6.35), and this is presented below.

There is a more general difficulty with many of the procedures developed so far in this chapter and in Chapters 4 and 5, and this relates to the fact that $\underline{\alpha}(r,r-u)$ is scalar diagonal: in short, the mover ratio matrix is independent of location. For complete sequence processes, this might be tenable as was argued in the second paper, but for *ad hoc* processes, a locationally independent mover matrix is too arbitrary. Of course, it is possible to extend the pseudo-dynamic model to deal with non-scalar diagonal mover matrices but the recurrence relations describing the movers and stayers no longer hold. To be consistent, each mover ratio must be applied to the original input $\Delta^*\underline{s}(0)$, and the sequence recomputed from there to the point where $\underline{\alpha}(r,r-u)$ is fixed. The storage problem

for these recomputed streams is horrific, and it illustrates that the need to be consistent in terms of the input-output sequences characterising the model, leads to computational and logical difficulties which limit the model.

This point can best be made pertinent in connection with the single mover stream model of this section. To ensure that equation (6.35) is correct for a non-scalar diagonal mover ratio matrix, the last term must be replaced by

$$\sum_{\tau=t-T}^{r-1} \Delta^*\underline{s}(\tau) - \sum_{\tau=t-T}^{r-2} \Delta^*\underline{s}(\tau)\underline{\alpha}(\tau+2,t-T+1)\underline{A}(\tau+2)\underline{B}(\tau+2)$$

$$- \Delta^*\underline{s}(0)\underline{\alpha}(t-T+1,t-T+1)\underline{A}(t-T+1)\underline{B}(t-T+1)$$

which clearly involves another stream of computation. In the model developed in Chapter 5 for the Reading region, there was no need to make $\underline{\alpha}(r,r-u)$ locationally dependent due to the fact that the activity was being relocated through a complete sequence. But here, a single mover stream model is to be developed and thus it would be desirable if such a model incorporated mover ratio matrices which depend upon location.

The basic idea on which a pseudo-dynamic model can be designed which meets the conditions mentioned above, involves the aggregation of the mover and new change sequences into one. As movers for the next time period are evaluated as a proportion of the movers and new change characterising the present period, it is clear that the mover ratio can never exceed unity. Moreover, this logic enables stayers to be computed directly and in fact, after the first time period, the new

change sequence is no longer distinguishable from the movers and stayers. Figure 6.2(c) makes this idea visually explicit and it is clear that although movers and the previous stayers together generate new change, the two sequences are aggregated in a manner which makes the computation of new change in the previous sense, laborious and unnecessary.

No longer is it possible to specify the period of simulation with this model, and although the life of the sequence is still T units, the simulation starting at t-T would end no earlier than t+1 and no later than t+T+1. In other words, because the mover and new change sequences are not strictly separated, the life of the combined sequence can only be found by simulation. In the development of this model, it is assumed that $\underline{A}(r) = \underline{\tilde{A}}(r)$ and $\underline{B}(r) = \underline{\tilde{B}}(r)$ which is the assumption used by Baxter and Williams (1975) in their model, and note that as the mover ratio matrix $\underline{\alpha}(r,t-T+1)$ is always dependent on t-T+1, this time script is omitted.

The new model can now be presented. The state equations for population and employment are quite simple for they are composed mainly of stayers due to the fact that movers and new change input to any one time period generate stayers at the end of the period, and more movers and new change for the next period. Then

$$\underline{p}(r) = \sum_{\tau=t-T+1}^{r} \Delta\underline{p}^{s}(\tau), \tag{6.36}$$

$$\underline{e}(r) = \Delta^{*}\underline{s}(0) + \sum_{\tau=t-T+1}^{r} \Delta\underline{s}^{s}(\tau), \tag{6.37}$$

where $\Delta\underline{p}^{s}(\tau)$ and $\Delta\underline{s}^{s}(\tau)$ are vectors of population and employment stayers generated at time $\tau$. In recursive form, the model can be written as

follows. Define $\Delta\hat{\underline{p}}(r)$ and $\Delta\hat{\underline{s}}(r)$ as vectors of population and employment change before movers have been evaluated by reference to constraints

$$\Delta\hat{\underline{p}}(r) = [\Delta\underline{s}^s(r-1) + \Delta\underline{s}^m(r-1)]\underline{A}(r), \tag{6.38}$$

$$\Delta\hat{\underline{s}}(r) = \Delta\hat{\underline{p}}(r)\underline{B}(r). \tag{6.39}$$

$\Delta\underline{s}^m(r-1)$ are the employment movers calculated in the previous time period.

At this point, $\Delta\hat{\underline{p}}(r)$ and $\Delta\hat{\underline{s}}(r)$ are used to test against the constraints $\underline{c}^p(r)$ and $\underline{c}^e(r)$, and some scheme adapted from equations (6.21) to (6.30) is used to calculate $\underline{\alpha}(r)$. Movers $\Delta\underline{s}^m(r)$ can now be calculated as

$$\Delta\underline{s}^m(r) = [\Delta\underline{s}^s(r-1) + \Delta\underline{s}^m(r-1)]\underline{\alpha}(r), \tag{6.40}$$

and the stayers can be found by allocating a proportion of the input as follows

$$\Delta\underline{p}^s(r) = [\Delta\underline{s}^s(r-1) + \Delta\underline{s}^m(r-1)][\underline{I}-\underline{\alpha}(r)]\underline{A}(r), \tag{6.41}$$

$$\Delta\underline{s}^s(r) = \Delta\underline{p}^s(r)\underline{B}(r). \tag{6.42}$$

The algorithm in equations (6.38) to (6.42) is operated until some limit defining the combined new change and mover sequences' life is reached. Note that the range for r begins at r=t-T and that $\Delta\underline{s}^s(t-T-1) = \Delta^*\underline{s}(0)$ and $\Delta\underline{s}^m(t-T-1) = \underline{0}$.

It is possible to derive detailed forms of state equation by successive substitution from equations (6.40) and (6.41) into themselves, and then substitution into equations (6.36) and (6.37). But the resulting forms are extremely complicated although quite regular, and nothing is gained

by this apart from a demonstration that the ultimate state of the system depends on a successive modulation of the mover and stayer ratios. The model however is not entirely complete for it has only been presented here in aggregative terms. It is necessary to present the model in the form in which it is computed, that is, in a non-matrix form for only then does the method for handling locational constraints become clear. Furthermore, as the emphasis now shifts to the calibration problem, this type of matrix notation becomes unwieldy for the focus is on a small number of system parameters and constraints rather than a large set of zones. Although it would be possible to present the rest of the chapter in matrix notation, it is preferable to present it in more usual terms so that relationships to other work concerning spatial interaction models is apparent.

COMPUTING THE SIMPLIFIED MODEL: THE DISTRIBUTION SUBMODELS.

It is necessary to assemble the general algorithm for the model in two main stages. First, the submodels which form the elements of the distribution matrices $\underline{A}(r)$ and $\underline{B}(r)$ will be outlined and these models which are based on interaction models of the gravity type are identified with the parameters of the general model to be determined through calibration. Then these submodels are embedded into the model's main equations which are based on equations (6.36) to (6.42), and this makes possible a detailed presentation of the algorithm used to effect locational constraints. Once this has been achieved, the general model can be considered ready for calibration and this involves setting this algorithm within a wider algorithm for calibration which is developed in the next chapter.

Associated with the distribution matrices $\underline{T}(r)$ and $\underline{S}(r)$ are probability distributions which show the probability of an interaction originating in the set of zones whose activity distribution is known, and terminating in the set whose activity distribution it is required to predict. In tracing the sequence from origin to destination which becomes a new origin and so on, the interaction associated with $\underline{T}(r)$ is indexed by origin i, destination j, and that associated $\underline{S}(r)$ by origin j (which is the same as the set of destinations {j}),destination k. However, because only two distributions of activity are involved, the j index refers to population in zones and the i, and k to employment. Note that it is assumed that the set of zones subscripted by i is identical to that subscripted by k.

Then the probability of working in i and living in j is defined as $p_{ij}(r)$ and the distribution is normalised so that

$$\sum_{i}\sum_{j} p_{ij}(r) = 1.$$

The marginal probability $t_i(r-1)$ is known from the previous time period and $p_{ij}(r)$ is also defined so that

$$\sum_{j} p_{ij}(r) = t_i(r-1).$$

The other set of marginal probabilities on destination locations $s_j(r)$ is to be predicted from

$$\sum_{i} p_{ij}(r) = s_j(r),$$

and it is also clear that

$$\sum_{i} t_i(r-1) = \sum_{j} s_j(r) = 1.$$

However the elements of the matrix $\underline{T}(r)$ are conditional probabilities defined with respect to the known marginal probabilities of the origins, that is

$$t_{ij}(r) = \frac{p_{ij}(r)}{t_i(r-1)} , \tag{6.43}$$

from which it is clear that the probability $p_{ij}(r)$ is calculated as

$$p_{ij}(r) = t_i(r-1)t_{ij}(r). \tag{6.44}$$

Note that $t_{ij}(r)$ sums to unity over j and this is sufficient to ensure that $\underline{T}(r)$ is row stochastic.

Exactly the same logic can be used to develop the probability $q_{jk}(r)$ associated with $\underline{S}(r)$. This probability is normalised so that

$$\sum_{jk}\sum q_{jk}(r) = 1,$$

and the associated marginal (origin and destination) probabilities are defined as

$$\sum_k q_{jk}(r) = s_j(r), \quad \text{and}$$

$$\sum_j q_{jk}(r) = t_k(r),$$

with the obvious normalisation

$$\sum_j s_j(r) = \sum_k t_k(r) = 1.$$

The conditional probability $s_{jk}(r)$ is calculated as

$$s_{jk}(r) = \frac{q_{jk}(r)}{s_j(r)} \quad , \tag{6.45}$$

from which the probability $q_{jk}(r)$ is easily determined using the probability rule for independent events

$$q_{jk}(r) = s_j(r)s_{jk}(r). \tag{6.46}$$

Note that $s_{jk}(r)$ sums to unity over k and this makes $\underline{S}(r)$ row stochastic, and that the distribution $s_j(r)$ must be known before $s_{jk}(r)$ can be calculated.

In fact, gravity type interaction models are postulated using the information-minimising framework outlined in Chapters 3 and 5 and these models determine the conditional probabilities $t_{ij}(r)$ and $s_{jk}(r)$. The marginal distributions depend upon the dynamic processes of the model itself and this can easily be seen from the recurrence which is implied in the following equations. Using equation (6.44) and the appropriate form for $t_{ij}(r)$, $p_{ij}(r)$ is defined as

$$p_{ij}(r) = t_i(r-1) \frac{q_{jk}(r-1)\exp\{-\mu_1(r)c_{ij}\}}{\sum_j q_{jk}(r-1)\exp\{-\mu_1(r)c_{ij}\}} \quad , \quad k=i, \tag{6.47}$$

and the marginal probability is calculated as

$$s_j(r) = \sum_i p_{ij}(r). \tag{6.48}$$

In equation (6.47), $\mu_1(r)$ is a parameter controlling the amount of interaction generated by the model and $c_{ij}$ is the generalised cost of travel between i and j, typically time-distance.

This submodel is derived using first order information-minimising, and

a similar form is postulated for $q_{jk}(r)$. Then

$$q_{jk}(r) = s_j(r) \frac{p_{ij}(r)\exp\{-\mu_2(r)c_{jk}\}}{\sum_{i,k} p_{ij}(r)\exp\{-\mu_2(r)c_{jk}\}}, \quad k=i, \tag{6.49}$$

and the probability of locating in k is thus calculated as

$$t_k(r) = \sum_j q_{jk}(r). \tag{6.50}$$

The way in which the submodels are operated is fairly obvious from the sequence given in equations (6.47) to (6.50). As the sets of zones {i} and {k} are identical, the procedure requires an initial set of probabilities $\{t_i(0)\}$ and by substitution of $t_k(r)$ from equation (6.50) for $t_i(r-1)$, k=i, in equation (6.47), a recurrence procedure through time is defined. This is a first order Markovian scheme with non-stationary transition probability matrices $\underline{T}(r)$ and $\underline{S}(r)$, and the absolute distributions of activity in employment and population zones are calculated when this locational scheme is incorporated into the multiplier sequences presented in previous sections.

The submodel forms in equations (6.47) and (6.49) can be derived in a manner similar to that used in Chapter 5 to derive equations (5.38) and (5.43). In fact, the whole set of equations from (5.37) to (5.44) is relevant in deriving (6.47) to (6.49) with equations (5.37) and (5.42) replaced by

$$\sum_{ij}\sum p_{ij}(r)c_{ij} = \bar{C}, \text{ and} \tag{6.51}$$

$$\sum_{jk}\sum q_{jk}(r)c_{jk} = \bar{S}, \tag{6.52}$$

respectively. Here it is explicitly assumed that the mean amount of

interaction generated by the workplace-home interaction called $\bar{C}$ and that generated by the home-service demand function called $\bar{S}$, are constant over time. This is a reasonable assumption given the nature of the pseudo-dynamic model, and as such, $\bar{C}$ and $\bar{S}$ represent fixed targets which the system must meet by the end of the simulation. The parameters $\mu_1(r)$ and $\mu_2(r)$ control the mean amount of interaction predicted in each time period, and it is through manipulation of these parameters that the targets are reached.

Equations (6.51) and (6.52) must therefore be solved for $\mu_1(r)$ and $\mu_2(r)$ at each time period r and thus the calibration problem becomes dynamic. Indeed, the strategy for determining trajectories for $\mu_1(r)$ and $\mu_2(r)$ is the central task of the next chapter and will be discussed at length in the sequel. Before the submodels are embedded into the main model, it is worth rewriting the trip length equations (6.51) and (6.52) in terms of marginal and conditional probabilities. Then substituting for $p_{ij}(r)$ and $q_{jk}(r)$ from equations (6.44) and (6.46) into (6.51) and (6.52) respectively gives

$$\bar{C} = \sum_i t_i(r-1)\sum_j t_{ij}(r)c_{ij} = \sum_i t_i(r-1)\bar{c}_i(r), \quad \text{and} \tag{6.53}$$

$$\bar{S} = \sum_j s_j(r)\sum_k s_{jk}(r)c_{jk} = \sum_j s_j(r)\bar{s}_j(r). \tag{6.54}$$

$\bar{c}_i(r)$ and $\bar{s}_j(r)$ are the mean zonal trip lengths which are weighted according to the zonal distribution of activity in the calculation of the mean system trip lengths $\bar{C}$ and $\bar{S}$. This is a useful interpretation which will be developed later when the effect of the multiplier sequences on the mean trip lengths is examined.

AN ALGORITHM FOR THE SIMPLIFIED MODEL.

The main algorithm which follows equations (6.36) to (6.42) can now be developed, thus demonstrating the way in which the locational submodels are embedded into the multiplier process, and the way in which constraints are handled. First, the population model is developed and the mover ratio $\underline{\alpha}(r)$ determined according to constraint violations in residential zones. The population stayers are then computed and used as an input to the employment submodels. This output is then evaluated against constraints and the mover ratio is further adjusted on this basis. The structure of the model requires that the population stayers be readjusted if employment constraints have been violated. The procedure presented below for one time period or model iteration involves the computation of a series of intermediate values for certain variables: such a variable is as defined previously but distinguished by the use of the circumflex ^ for a first value and the double circumflex for a second value $\hat{\hat{}}$.

From the previously predicted distributions of employment stayers $\{\Delta S_i^s(r-1)\}$ and movers $\{\Delta S_i^m(r-1)\}$, the first estimate of work trips $\Delta\hat{T}_{ij}(r)$ is calculated from

$$\Delta\hat{T}_{ij}(r) = [\Delta S_i^s(r-1) + \Delta S_i^m(r-1)]t_{ij}(r), \tag{6.55}$$

and a first estimate of the population $\Delta\hat{P}_j(r)$ is derived by applying the activity rate $\lambda$ to the sum of trips terminating in j

$$\Delta\hat{P}_j(r) = \lambda\sum_i\Delta\hat{T}_{ij}(r). \tag{6.56}$$

At this point, the constraints on population are checked: if

$$[\Delta\hat{P}_j(r) + P_j^s(r-1)] \geq C_j^p(r), \quad j \text{ is assigned to } Z_p, \tag{6.57}$$

and the surplus is computed from

$$\Delta_j^p(r) = \Delta\hat{P}_j(r) + P_j^s(r-1) - C_j^p(r), \qquad j\varepsilon Z_p. \tag{6.58}$$

The proportion of the input employment used to generate change in this time period which is associated with this surplus is calculated for each zone i by finding the proportion of trips $\Delta T_{ij}(r)$ forming $\Delta_j^p(r)$ and reallocating back to each employment origin. Then

$$\sigma_i(r) = \theta_i(r) \sum_{j\varepsilon Z_p} \frac{\Delta_j^p(r)}{\Delta\hat{P}_j(r)} \Delta\hat{T}_{ij}(r), \tag{6.59}$$

where the coefficient $\theta_i(r)$ is defined as

$$\theta_i(r) = 1/[\Delta S_i^s(r-1) + \Delta S_i^m(r-1)]. \tag{6.60}$$

It is now necessary to adjust the trips and the population to account for these constraint violations, and new intermediate variables $\Delta\hat{\hat{T}}_{ij}(r)$ and $\Delta\hat{\hat{P}}_j(r)$ are computed

$$\Delta\hat{\hat{T}}_{ij}(r) = \Delta\hat{T}_{ij}(r)[1 - \frac{\Delta_j^p(r)}{\Delta\hat{P}_j(r)}], \quad j\varepsilon Z_p, \quad \text{and} \tag{6.61}$$

$$\Delta\hat{\hat{P}}_j(r) = \Delta\hat{P}_j(r) - \Delta_j^p(r), \qquad j\varepsilon Z_p. \tag{6.62}$$

Note here that $\sigma_i(r)$ is a first estimate for $\alpha_i(r)$ based only on the population sector and this must be further modified to account for any constraint violations due to the allocation of surplus employment.

An analogous procedure is used to allocate employment. First the demands for services by the population $\Delta\hat{\hat{P}}_j(r)$ are calculated as $\Delta\hat{S}_{jk}(r)$. Then

$$\Delta\hat{S}_{jk}(r) = \Delta\hat{\hat{P}}_j(r)s_{jk}(r), \tag{6.63}$$

and employment in k, $\Delta\hat{S}_k(r)$ is calculated by summing equation (6.63) over j and scaling by the population-serving ratio $\gamma$

$$\Delta\hat{S}_k(r) = \gamma\sum_j\Delta\hat{S}_{jk}(r). \tag{6.64}$$

Constraints on employment are now checked: if

$$[\Delta\hat{S}_k(r) + S_k^s(r-1)] \geq C_k^e(r), \text{ k is assigned to } Z_e, \tag{6.65}$$

and the surplus is computed from

$$\Delta_k^e(r) = \Delta\hat{S}_k(r) + S_k^s(r-1) - C_k^e(r), \quad k\varepsilon Z_e \quad . \tag{6.66}$$

The proportion of population associated with this surplus $\rho_j(r)$ is calculated as

$$\rho_j(r) = \phi_j(r) \sum_{k\varepsilon Z_e} \frac{\Delta_k^e(r)}{\Delta\hat{S}_k(r)} \Delta\hat{S}_{jk}(r), \tag{6.67}$$

where $\phi_j(r)$ is defined as

$$\phi_j(r) = 1/\Delta\hat{\hat{P}}_j(r). \tag{6.68}$$

At this point, it is necessary to adjust the trips $\Delta\hat{S}_{jk}(r)$ and the employment $\Delta\hat{S}_k(r)$ to account for constraint violations. Then

$$\Delta S_{jk}(r) = \Delta\hat{S}_{jk}(r)[1 - \frac{\Delta_k^e(r)}{\Delta\hat{S}_k(r)}], \quad k\varepsilon Z_e \text{ , and} \tag{6.69}$$

$$\Delta S_k^s(r) = \Delta\hat{S}_k(r) - \Delta_k^e(r), \quad k\varepsilon Z_e \text{ .} \tag{6.70}$$

However as the model is structured in terms of a single input-employment, it is necessary to transform the surplus population associated with $\Delta_k^e(r)$

into surplus employment based on the input to time period r. Thus the final stayer population $\Delta P_j^s(r)$ is calculated as

$$\Delta P_j^s(r) = \Delta \hat{\hat{P}}_j(r)[1 - \rho_j(r)], \tag{6.71}$$

and work trips as

$$\Delta T_{ij}(r) = \Delta \hat{\hat{T}}_{ij}(r)\ [1 - \rho_j(r)]. \tag{6.72}$$

The final mover ratio $\alpha_i(r)$ is updated according to the new surplus associated with $\rho_j(r)$ and thus

$$\alpha_i(r) = \sigma_i(r) + \phi_i(r)\sum_j \rho_j(r)\Delta \hat{\hat{T}}_{ij}(r), \tag{6.73}$$

and the movers $\Delta S_i^m(r)$ to be allocated in the next time period are computed from

$$\Delta S_i^m(r) = [\Delta S_i^s(r-1) + \Delta S_i^m(r-1)]\alpha_i(r). \tag{6.74}$$

Note that the stayers are computed as the process of evaluating the movers is accomplished within the time period, and that the ratio $\alpha_i(r)$ is separable into a component associated with population surplus and one associated with employment surplus.

The structure of the general model is presented as a flow diagram in Figure 6.3 and this diagram will be used later as part of the more general flow chart developed to illustrate the calibration procedure. At this point, a number of quantities relating to the state of the system at time r must be calculated. Cumulative totals of trips and activities are calculated as follows:

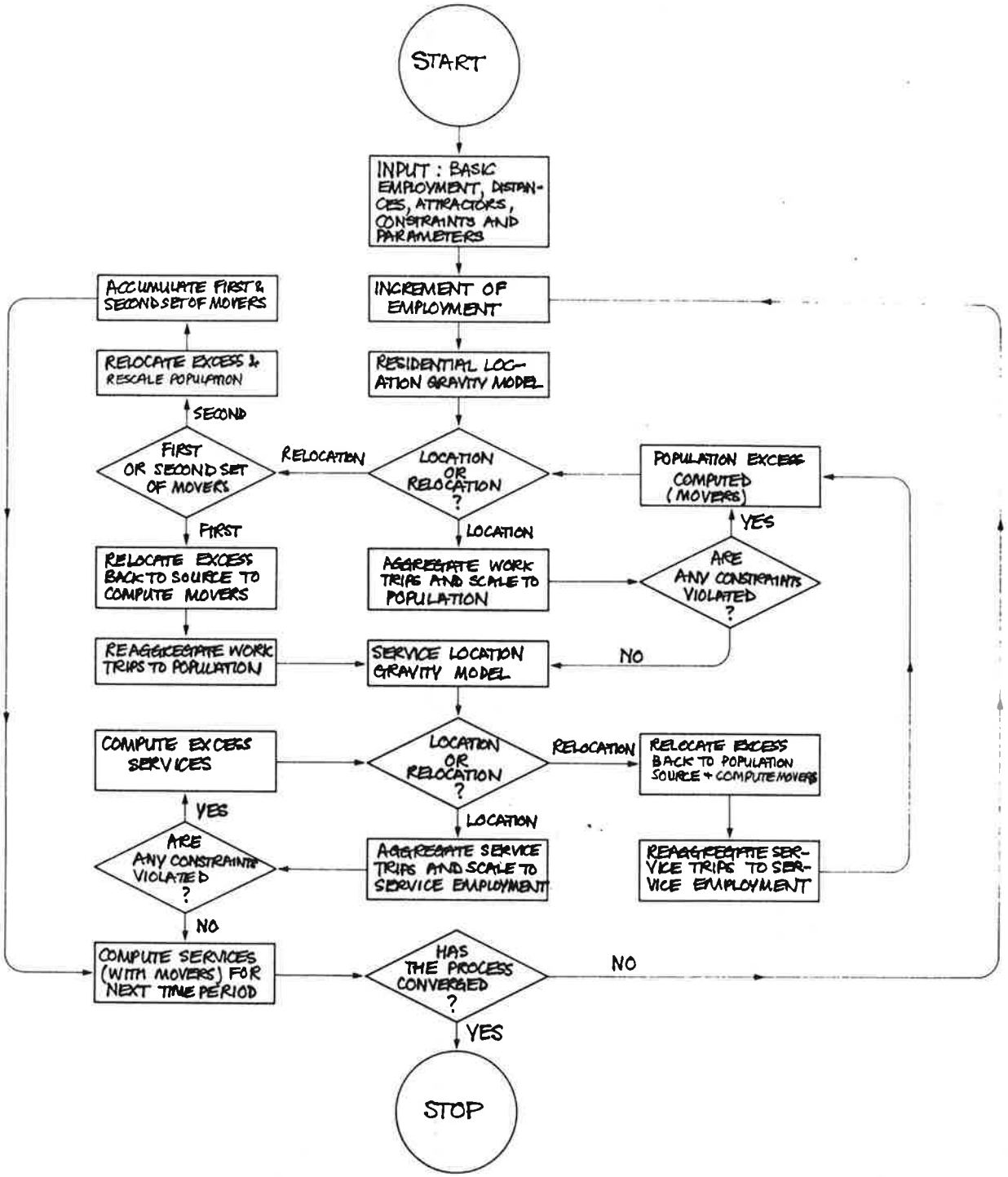



Figure 6.3: Sequence of Operations in the Simplified Pseudo-Dynamic Model.

$$\left.\begin{aligned} T_{ij}(r) &= T_{ij}(r-1) + \Delta T_{ij}(r), \\ P_j(r) &= P_j(r-1) + \Delta P_j^s(r), \\ S_{jk}(r) &= S_{jk}(r-1) + \Delta S_{jk}(r), \qquad \text{and} \\ S_k(r) &= S_k(r-1) + \Delta S_k^s(r). \end{aligned}\right\} \quad (6.75)$$

Of particular importance however are the mean trip lengths which indicate the dimensional fit of the model to reality, and these are of use in guiding the calibration procedure developed below.

The mean work trip length associated with the change in time period r is called $\Delta\bar{C}(r)$ and is defined as

$$\begin{aligned} \Delta\bar{C}(r) &= \sum_{ij}\sum \Delta T_{ij}(r)c_{ij} / \sum_{ij}\sum \Delta T_{ij}(r), \\ &= \sum_i t_i(r-1) \sum_j t_{ij}(r)c_{ij}, \end{aligned} \quad (6.76)$$

where $t_i(r-1)$ is the distribution of input employment used to generate $\{\Delta P_j^s(r)\}$ and calculated as

$$t_i(r-1) = \frac{[\Delta S_i^s(r-1) + \Delta S_i^m(r-1)][1-\alpha_i(r)]}{\sum_i [\Delta S_i^s(r-1) + \Delta S_i^m(r-1)][1-\alpha_i(r)]} \quad . \quad (6.77)$$

The service demand mean trip length $\Delta\bar{S}(r)$ is calculated in an analogous way

$$\begin{aligned} \Delta\bar{S}(r) &= \sum_{jk}\sum \Delta S_{jk}(r)c_{jk} / \sum_{jk}\sum \Delta S_{jk}(r), \\ &= \sum_j s_j(r) \sum_k s_{jk}(r)c_{jk}, \end{aligned} \quad (6.78)$$

where $s_j(r)$ is defined as

$$s_j(r) = \frac{\Delta P_j^S(r)}{\sum_j \Delta P_j^S(r)} \tag{6.79}$$

The cumulative trip lengths $\bar{C}(r)$ and $\bar{S}(r)$ can be calculated by substituting $T_{ij}(r)$ and $S_{jk}(r)$ from equations (6.75) into the first term of equations (6.76) and (6.78) respectively, although there is a more fundamental relationship between the previous trip lengths and the change. This will be detailed later as it is central to the idea of guiding the system towards the fixed targets $\bar{C}$ and $\bar{S}$.

The final stages of computation in time period r consist of setting up the matrices $\{t_{ij}(r+1)\}$ and $\{s_{jk}(r+1)\}$ for the next time period. As the submodels which structure these matrices are based on first order lags in the probability distributions, $p_{ij}(r)$ and $q_{jk}(r)$ must be calculated

$$p_{ij}(r) = \frac{\Delta T_{ij}(r)}{\sum_{ij}\sum \Delta T_{ij}(r)} , \qquad \text{and}$$

$$q_{jk}(r) = \frac{\Delta S_{jk}(r)}{\sum_{jk}\sum \Delta S_{jk}(r)} .$$

The parameters $\mu_1(r+1)$ and $\mu_2(r+1)$ are fixed exogenously or according to the calibration algorithm and $t_{ij}(r+1)$ and $s_{jk}(r+1)$ are then computed using equations (6.47) and (6.49). However, if constraints have been violated in time period r, the matrices $\underline{T}(r)$ and $\underline{S}(r)$ must be renormalised so that

$$t_{ij}(r+1) = 0, \quad j\varepsilon Z_p, \quad \text{and } s_{jk}(r+1) = 0, \quad k\varepsilon Z_e.$$

$\Delta S_i^s(r)$ and $\Delta S_i^m(r)$ are now substituted into equation (6.55) and equations (6.55) to (6.79) are reiterated until the life of the process is complete. The life of the process may be set arbitrarily at T+1 time periods, or as in this case, it is determined during the simulation according to the cut-off limit $\varepsilon$. Then if

$$\frac{(1-\gamma\lambda)\sum_i \Delta S_i^s(r)}{\sum_i \Delta^* S_i(0)} \leq \varepsilon \quad ,$$

the simulation is terminated. In this model, $\varepsilon$ was set equal to 0.01 although there are several ways of approximating convergence to such a limit which reduce computation time (Batty, 1976). At this point, all the elements have been presented which enable a comprehensive discussion of the calibration problem to take place, and this will be begun in the next chapter.

CONCLUSIONS.

At this point, we are about halfway through this thesis and it is worth reflecting on progress so far. In effect, the internal structure of the conventional urban model due to Lowry (1964) has been elaborated through its multiplier relations. These relations embody a type of pseudo-time, and in their elaborated form, such models have been called pseudo-dynamic. This form of model was developed in Chapter 3, and since then many variants of this model have been presented. In Chapter 4, these models were treated as mechanisms for enabling locational constraints to be handled and in Chapter 5, some examples were given. In this chapter another variant of the pseudo-dynamic model has been outlined suitable not only for embodying locational constraints but also for enabling

spatial interaction submodels to be calibrated. The algorithm outlined here will be embedded in wider process of calibration in Chapter 7 which in turn will lead in Chapters 8 and 9 to more efficient algorithms combining model solution with locational constraints and spatial interaction calibration.

# CHAPTER 7.

# AN ALGORITHM FOR ADAPTIVE CALIBRATION.

The conventional procedure in estimating the parameters of general urban models of the Lowry type involves embedding the model into some wider iterative process in which parameter values are optimised. For example, many *ad hoc* schemes based on simple iteration exist (see Batty, 1976) while more recently, similar sorts of models have been formulated as constrained optimisation problems and solved accordingly (Wilson, Coelho, Macgill and Williams, 1981). However in these cases either the model solution is embedded into a calibration process or solution and calibration are achieved simultaneously. The ideas developed in previous chapters suggest that both constraint and calibration procedures can be embedded into model solution mechanisms, if such mechanisms have a tractable and sequential form. In this case, this form is essentially that of the pseudo-dynamic process and in the last chapter, an algorithm which involved exploiting this process was developed in the context of locational constraints. In this chapter the algorithm will be extended by embedding within it the mechanisms required to calibrate the model.

## THE DYNAMIC CALIBRATION PROBLEM.

The calibration problem has already been defined and posed informally in

Chapter 6 as one in which it is required to find a sequence of parameter values $\mu_1(r)$, $\mu_2(r)$ which ensure that equations (6.51) and (6.52) are solved for every time period r. In short, this implies that $\Delta\bar{C}(r) = \bar{C}$, and $\Delta\bar{S}(r) = \bar{S}$, $\forall r$, and as such, the problem can be seen as one of optimal control in which it is required to optimise some function of the difference between the predicted and intended mean trip lengths through the simulation period. In this sense, the parameters $\mu_1(r)$ and $\mu_2(r)$ act as the instruments of the process; $\mu_1(r)$ and $\mu_2(r)$ are independent of their previous values, that is , no autocorrelation is implied by this process. Classical methods of control, however, are concerned with deriving recursive procedures which successively update the parameters and thus establish a sequence which is efficient in some sense. Such procedures tend to be appropriate to well-defined and mathematically tractable linear state equations in which the optimisation can be accomplished using some linear feedback control rules which determine future values of the parameters of policy-control variables. For instance, the use of the Riccati equation for simple linear systems is a well-known means for optimising such a system in terms of a quadratic welfare function (Chow, 1975).

The objective function adopted here is based on the squared deviations between predicted and intended trip lengths. It is required to find $\mu_1(r)$ and $\mu_2(r)$ so that $[\Delta\bar{C}(r)-C]^2 = 0$ and $[\Delta\bar{S}(r)-\bar{S}]^2 = 0$, and the composite sum of squares function to be minimised is defined as

$$\Delta Z(r) = [\Delta\bar{C}(r)-\bar{C}]^2 + [\Delta\bar{S}(r)-\bar{S}]^2, \qquad (7.1)$$

which is clearly equal to zero for a solution to equations (6.51) and (6.52) to exist. In fact, it is possible to set up a Lagrangean based on the function in equation (7.1) using the state equations as constraints

and summing these objective functions and their constraints over r. The usual conditions for a minimum hold, and in this case, the solution would imply that equation (7.1) be solved exactly in each time period.

Such an exercise would merely show that the model would need to be solved recursively from the initial time period to the end of the simulation, and that equations (6.51) and (6.52) would need to be solved using a procedure such as the Newton-Raphson method in each time period. This is the obvious method implied in the outline of the model so far. Suffice it to say that the model is sufficiently nonlinear to hinder any more elegant solution procedure which improves on the recursive structure presented above. Nevertheless, one point does emerge from this argument and that is that the sequential nature of the general model makes possible a somewhat faster method of solution which avoids the simultaneous structure of the fully static model. This will be elaborated below.

Although it has been suggested that $\Delta Z(r)$ be minimised in each time period, the central interest in this process of optimisation relates to the notion that the sum of the objectives over the whole simulation must be minimised. That is

$$\sum_r \Delta Z(r) = 0, \tag{7.2}$$

and this implies that by the end of the simulation, the cumulative trip length $\bar{C}(r)$ and $\bar{S}(r)$ are equal to their intended values, $\bar{C}$ and $\bar{S}$. In other words, as long as the whole model reproduces the intended values, the process through which these values are reached within the model is of no significance. This is a fairly reasonable supposition because the intended trip lengths are specified for the whole process anyway due to the way in which they are observed. A time series for these trip lengths

is simply not relevant because the dynamic process is an approximation used to generate a static situation.

If equation (7.2) were the objective, then each $\Delta Z(r)$ would not necessarily be a minimum although the sum of these values may be. The advantage of relaxing the optimisation in this way would enable an approach to be pursued in which suboptimisation of the objective in any particular time period could be allowed in order to optimise the overall objective, and this could be used to speed up the process. For example, the parameters $\mu_1(r)$ and $\mu_2(r)$ would be chosen so that the model eventually moved towards the global optima through its pseudo-dynamic process, and in each time period, $\mu_1(r)$ and $\mu_2(r)$ would be adjusted in the effort to get nearer to the target.

There is another reason why the mean trip length statistics do not have any real significance in each time period. It can easily be demonstrated that for constant $\mu_1$ and $\mu_2$, the values of $\Delta\bar{C}(r)$ and $\Delta\bar{S}(r)$ predicted by the model vary. Thus in the case of the Reading model developed in Chapter 5, although the model was calibrated statically in that parameters $\mu_1$ and $\mu_2$ were found so that $\bar{C}(r)$ and $\bar{S}(r)$ predicted on the final (r'th) iteration of the model, met their intended values, the trip lengths predicted by the model changed in each time period. Indeed, it is quite easy to show this in both theoretical and practical ways.

For example, assume that the pseudo-dynamic model has no movers, that is, $\underline{\alpha}(r) = \underline{0}$, $\forall r$, that the trip distribution matrices $\underline{T}(r) = \underline{T}$ and $\underline{S}(r) = \underline{S}$, and that the parameters $\mu_1(r) = \mu_1$, $\mu_2(r) = \mu_2$ and are fixed exogenously. Such a model is equivalent to the Garin-Harris version of Lowry's (1964) Pittsburgh model (Batty, 1976) but despite the constancy in input, the

trip lengths change through its pseudo-dynamic process due to the successive compounding of the matrices $\underline{T}$ and $\underline{S}$. In short, the change in trip lengths can be seen as a consequence of the fact that the allocation procedure in the model is a regular Markov process with a stationary transition probability matrix $\underline{T}\ \underline{S}$ and a unique steady state solution. An analysis of this version of the model is presented in Appendix 2 where it is clear that the recursive structure of the model is responsible for variation in trip lengths.

A practical demonstration of this arbitrary change in mean trip lengths is presented in Figures 7.1 (a) to (d) for the model of this and the previous chapter with and without constraints. The model presented in equations (6.55) to (6.79) has been run for a fixed set of parameters $\mu_1(r) = \mu_1$ and $\mu_2(r) = \mu_2$, and although it cannot be analysed as a simple Markovian process as is the $\underline{\alpha}=\underline{0}$ model in Appendix 2, the regular change in mean trip length is apparent. Figures 7.1 (a) and (b) show the change in predicted trip lengths $\Delta\bar{C}(r)$, $\Delta\bar{S}(r)$, $\bar{C}(r)$ and $\bar{S}(r)$ for this model, and Figures 7.1 (c) and (d) the contribution of each time specific trip length to the total. The point of these theoretical and practical demonstrations of changes in the mean trip lengths without changes in the parameter values, is to show the arbitrariness of the assumption that there is a constant trip length for each time period.

This immediately raises the central issue on which the notion of adaptive calibration is based: because there is inevitable variation in the trip length due to the structure of the model,it is possible to accept and utilise this variation in homing in towards the ultimate targets as the simulation proceeds. Thus the trip length can be consciously varied in an

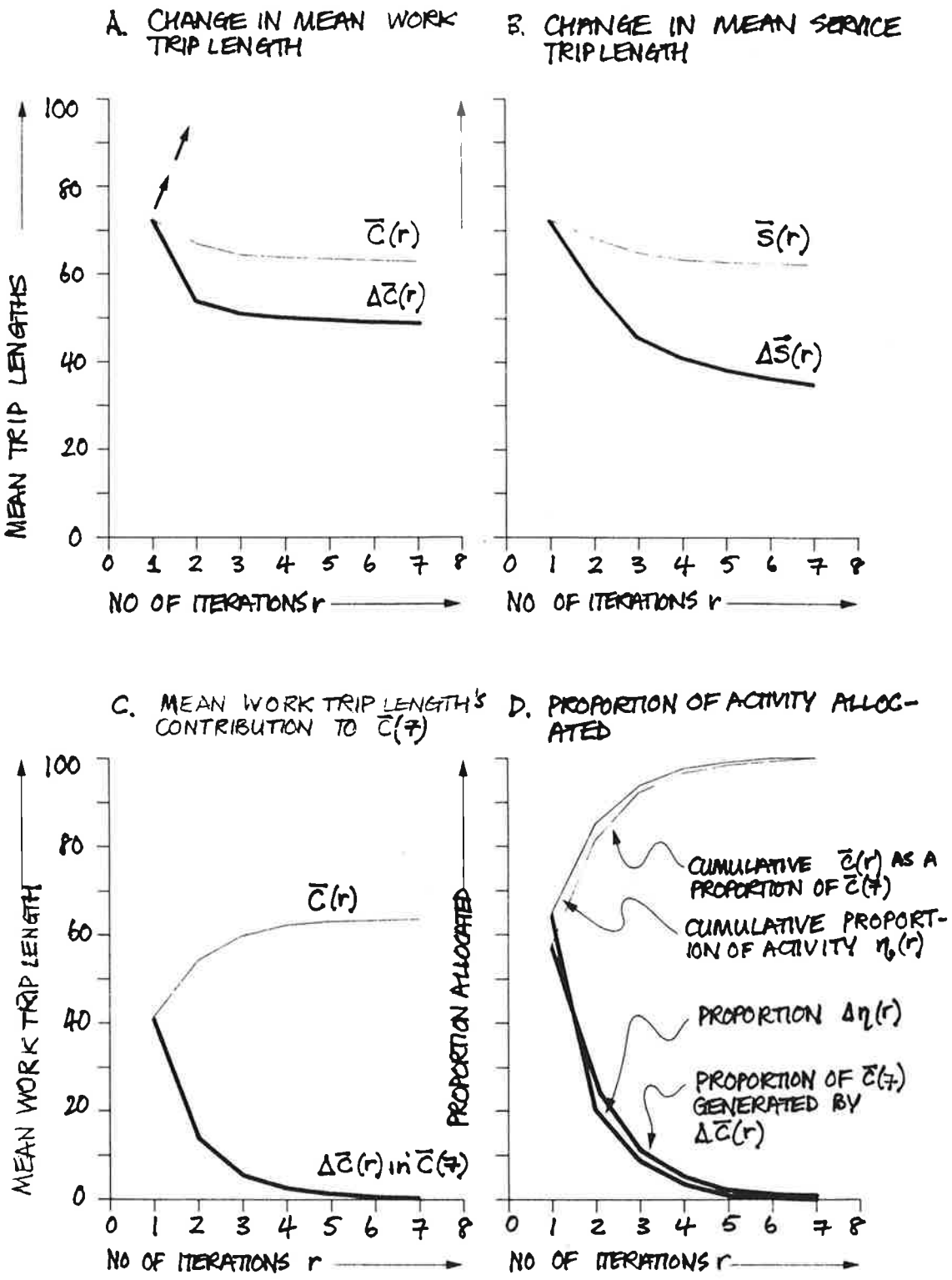



Figure 7.1: Changes in Trip Lengths through the Simulation Period.

effort to reach the global optima, and this represents support for the idea that $\mu_1(r)$ and $\mu_2(r)$ be varied in order that the optimisation take place. For example, in Figure 7.1(a) after the first iteration (time period) of the model in which $\Delta\bar{C}(1)$ has been predicted, the arrow points the direction to the value of $\Delta\bar{C}(2)$ which needs to be predicted thereafter if the system is to meet its intended value $\bar{C}$. Thus the algorithm to be suggested is based on the idea that the final targets are achieved by continually re-evaluating the time period targets to be met, and adjusting the values of the parameters to meet them.

A SKETCH OF THE ADAPTIVE SOLUTION PROCEDURE.

If the objective function were to be minimised in each time period, it would be necessary to solve the appropriate equations iteratively due to their intrinsic nonlinearity. The idea here is that the iterative structure of solving these equations is matched to the pseudo-dynamic structure of the model. Thus the objective function is never met in any one time period, but the overall objective is continually kept in mind, and the movement towards this overall objective is gradually attained by a sequence of partial solutions to the objective of each time period. Convergence would be guaranteed if the solution spaces for each time period objective were identical to one another but this is not the case. Because of the way in which activity is generated and allocated in the model, each trip distribution submodel is different from the same sub-model in a previous time period; thus a partial or full solution to a previous submodel in terms of its parameters, need not be useful in finding the parameters for the next time period.

However, it is hypothesised here that the solution spaces for each time

period are *sufficiently* similar to the previous time period for previous values of the parameters to be useful starting points for a partial solution to the set of parameters. The algorithm is thus designed on the notion of adapting, evolving or updating previous values of the parameters to meet the target trip length required for the next time period, in such a way that the trajectory of parameters ensures that the model meets its overall objectives: that is, that the intended mean trip lengths are met.

Several complications to this scheme are immediately apparent. Each trip length $\Delta\bar{C}(r)$ and $\Delta\bar{S}(r)$ makes less and less contribution to the cumulative trip lengths $\bar{C}(r)$ and $\bar{S}(r)$ due to the fact that less activity is allocated as the dynamic process works itself out. For example, in Figure 7.1 (d), 57% of activity is allocated in the first time period, 24% in the second and so on, and it is clear from Figure 7.1 (c) that the cumulative trip lengths get harder and harder to change. All things being equal, greater and greater changes in the target trip lengths are required to make further impacts on the cumulative trip length. If progress towards the intended target is too slow, then there may come a point in the simulation when the target trip lengths needed are impossible to meet for physical reasons. If the trip target became ridiculously excessive, for example, or fell towards zero, then this problem would emerge.

In short, there are bounds on what can be achieved in any time period and whether or not the simulation continues will depend on whether or not the trip targets are within these bounds. Moreover, the fact that the sequential process of the model affects the predicted trip lengths might also affect movement towards the targets. For example, in the model of this chapter, the trip lengths fall naturally through the time periods, and this exacerbates the fact that as time goes on, the intended

trip lengths get harder to meet.

Two elements of the algorithm are suggested by this discussion. First, there is the need to assess the trip targets required to ensure that the system meets its intended objectives by the end of the simulation. Because the model is based on an additive sequence of activity generation and allocation, all quantities associated with this sequence are additive. Thus as the cumulative trip length is known in any time period, and as the proportion of activity yet to be generated and allocated is known, it is a simple matter to calculate a target trip length needed to meet the intended (exogenous) target.

The second element of the algorithm involves the assessment of upper and lower bounds on the trip lengths feasible for the system. To calculate these bounds, it is possible to use Evans' (1973) results on the limiting forms of the gravity model, in which she showed that a maximum trip length was obtained when the parameter of the model tended to $-\infty$ and a minimum when the parameter tended to $+\infty$. The non-negativity properties of the model ensure that both these bounds are positive and Evans demonstrated that they were equivalent to the maximisation and minimisation of a linear objective function incorporating the generalised cost of travel subject to the normal origin and destination constraints on the model. In fact, the bounds can be obtained by solving the maximum and minimum problems associated with the transportation (linear programming) version of the gravity model. In the algorithm developed here, an approximation to these bounds is made to minimise computer time. This is elaborated in the sequel.

A third element of the algorithm is necessary if the trip targets are then found to be out of bounds. In such a case, the model associated with the time period in which the violation of the bounds occurred, must be rerun,

and an attempt made to calculate a target within bounds. This requires a means for readjusting the parameters for that time period and if after a certain number of trials, the bounds are still exceeded, it is assumed that there is no solution. The experimental work presented later is designed to explore and counter the circumstances surrounding such a possibility.

Having found a target within bounds, the fourth and final element of the algorithm involves finding a set of parameter values which will ensure that the model reaches or at least approaches the target in the following time period. This requires that the trip length equations associated with the targets be solved for $\mu_1(r)$ and $\mu_2(r)$, or that the sum of squares function based on the targets be minimised. In essence, the response surface associated with the solution space is approximated by a linear or quadratic surface and this enables the direction of the optimum to be established. For example, the Newton-Raphson procedure works on this idea. Clearly, for an exact solution, the surface must be continually approximated until the optimum is reached, but here it is necessary to explore the degree to which previous parameter values can be used as starting points in the approximation, and the degree to which an exact or approximate solution is necessary in the context of the overall simulation.

Finally, the solutions to the associated trip length least-squares normal equations is not simultaneous but sequential due to the dependence of activities in any one time period. These elements together with the original model given in the flow chart in Figure 6.3 are woven together in the algorithm presented in Figure 7.2. The general structure of the algorithm is clear from the above description and Figure 7.2 but it is now necessary to outline the elements in more detail before the experimental results are described.

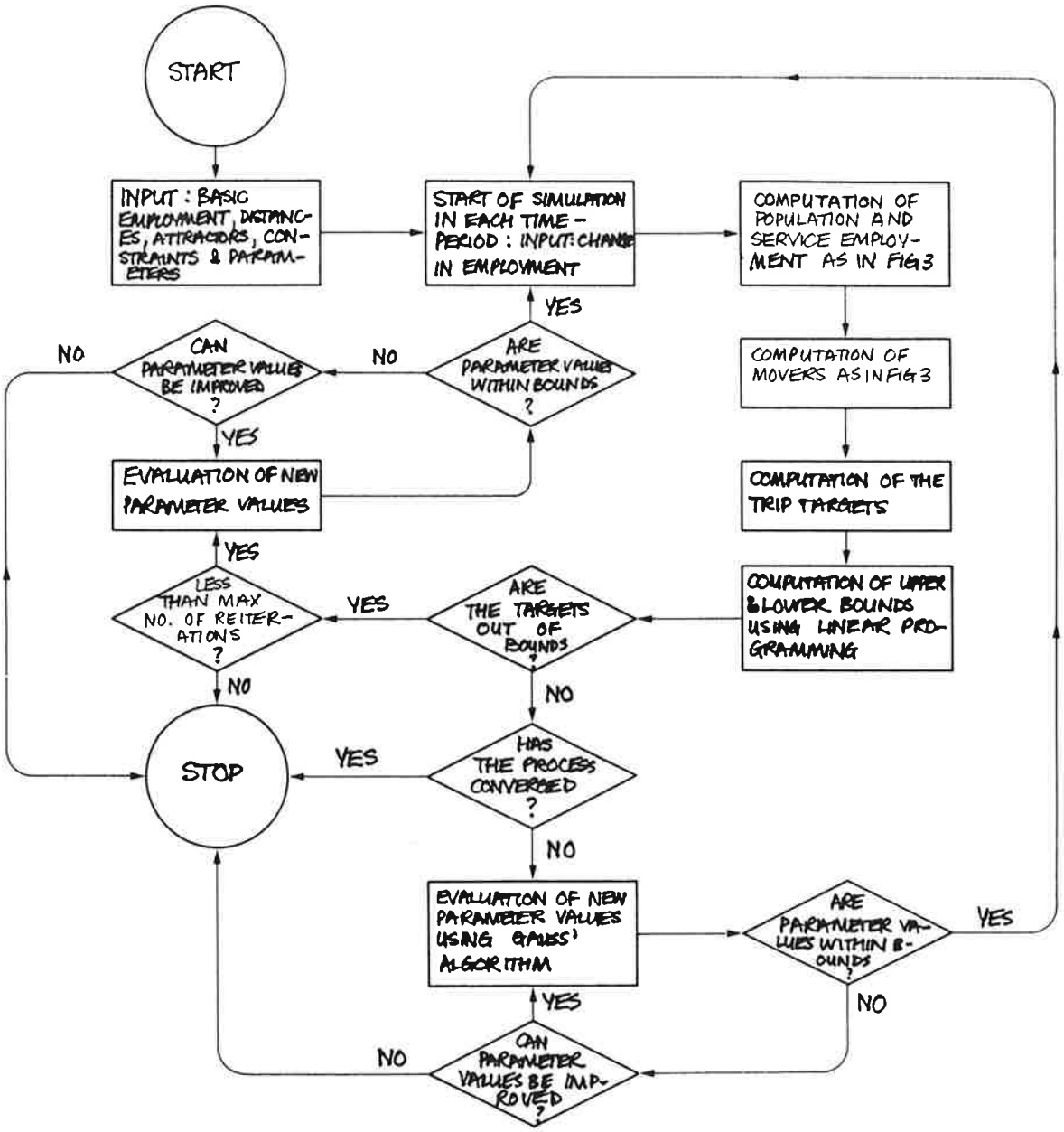



Figure 7.2: Elements in the Algorithm for Adaptive Calibration.

TRIP LENGTH TARGETS.

Assume that the model has just been run up to time period r and it is required to find the trip length targets for time period r+1 such that the intended target is met in all subsequent time periods. Then $\Delta\bar{C}(r)$, $\bar{C}(r)$, $\Delta\bar{S}(r)$ and $\bar{S}(r)$ can be computed directly, $\bar{C}$ and $\bar{S}$ are known a priori, and the proportion of activity generated so far is also known. Thus the amount of activity yet to be generated can be calculated and the amount to be allocated in the next time period can be approximated. The first stage in calculating the new trip length targets consists of finding a form for the proportion of activity generated, and re-expressing the trip lengths using this proportion. This enables an additive expression for the contribution of each trip length associated with each time period to be developed so that the precise relationship between each time period trip length and the cumulative statistic is determined.

As the employment and population state equations are related by simple scaling factors in an absolute sense, it is only necessary to develop a form for one set of trip lengths, say $\Delta\bar{C}(r)$ and $\bar{C}(r)$, for the other forms for $\Delta\bar{S}(r)$ and $\bar{S}(r)$ immediately follow by analogy. These analogies will be adopted here. Then the proportion of employment activity $\Delta\eta(r)$ associated with time period r, and taken as a proportion of the total employment E to be generated is given by

$$\Delta\eta(r) = \{\sum_i [\Delta S_i^s(r-1) + \Delta S_i^m(r-1)]\ [1-\alpha_i(r)]\}/E, \tag{7.3}$$

where E is calculated from the economic base equation

$$E = \sum_i \Delta^* S_i(0)\ (1-\gamma\lambda)^{-1} \quad . \tag{7.4}$$

Clearly the cumulative proportion of employment activity generated so far is

$$\eta(r) = \sum_{\tau=1}^{r} \Delta\eta(\tau), \tag{7.5}$$

and this proportion holds for population due to the fact that the scalar $\lambda$ connects total employment with population.

It is now possible to write the change in trips, $\Delta T_{ij}(r)$ and $\Delta S_{jk}(r)$, in terms of the proportion of activity generated $\Delta\eta(r)$. Using equations (6.77) and (7.3), the change in work trips $\Delta T_{ij}(r)$ can be written as

$$\Delta T_{ij}(r) = E\Delta\eta(r)t_i(r)t_{ij}(r), \tag{7.6}$$

and for service demands $\Delta S_{jk}(r)$, equations (6.79) and (7.3) yield

$$\Delta S_{jk}(r) = \lambda E\Delta\eta(r)s_j(r)s_{jk}(r). \tag{7.7}$$

The cumulative work trip length $\bar{C}(r)$ is given by the standard equation

$$\bar{C}(r) = \frac{\sum_{\tau=1}^{r} \sum_{ij}\sum \Delta T_{ij}(\tau)c_{ij}}{\sum_{\tau=1}^{r} \sum_{ij}\sum \Delta T_{ij}(\tau)}, \tag{7.8}$$

and substituting for $\Delta T_{ij}(\tau)$ from equation (7.6) gives the required form

$$\bar{C}(r) = \frac{E[\sum_{\tau=1}^{r} \Delta\eta(\tau)\sum_i t_i(r)\sum_j t_{ij}(r)c_{ij}]}{E\sum_{\tau=1}^{r} \Delta\eta(\tau)}$$

$$= \sum_{\tau=1}^{r} \Delta\eta(\tau)\Delta\bar{C}(\tau) / \sum_{\tau=1}^{r} \Delta\eta(\tau). \tag{7.9}$$

Exactly the the same procedure can be used to give the appropriate equation for $\bar{S}(r)$ which is stated as

$$\bar{S}(r) = \sum_{\tau=1}^{r} \Delta\eta(r)\Delta\bar{S}(r) / \sum_{\tau=1}^{r} \Delta\eta(\tau) \tag{7.10}$$

where it is clear that the cumulative trip length can be seen as a weighted average of the trip lengths associated with each time period.

From equations (7.9) and (7.10), it is immediately apparent that the trip lengths associated with later time periods have a lesser effect on the cumulative total than earlier trip lengths due to the fact that $\Delta\eta(r)$ converges as r increases. That is

$$\lim_{r\to\infty} \Delta\eta(r)\to 0, \qquad \text{and} \qquad \lim_{r\to\infty} \eta(r)\to 1,$$

and thus the mean work trip length in the limit, defined as $\bar{\bar{C}}$ is given as

$$\bar{\bar{C}} = \sum_{\tau=1}^{\infty} \Delta\eta(\tau)\Delta\bar{C}(\tau) \ . \tag{7.11}$$

From equation (7.11), it is obvious that $\bar{\bar{C}}$ can be separated into a component associated with the activity generated so far, and a component based on activity yet to be generated. Then

$$\bar{\bar{C}} = \sum_{\tau=1}^{r} \Delta\eta(\tau)\Delta\bar{C}(\tau) + \sum_{\tau=r+1}^{\infty} \Delta\eta(\tau)\Delta\bar{C}(\tau) \ ,$$

$$= \eta(r)\bar{C}(r) + [1-\eta(r)]\bar{\bar{C}}(r), \tag{7.12}$$

where $\bar{\bar{C}}(r)$ is the cumulative trip length associated with the simulation after time period r from r+1 to $\infty$. A similar equation can be developed for $\bar{\bar{S}}$ in terms of $\bar{S}(r)$ and $\bar{\bar{S}}(r)$. Note that equation (7.12) is also a weighted average at a higher level of temporal aggregation.

Equation (7.12) is the central equation in the derivation of trip length

targets for it implies that the final trip length is a function of the cumulative trip length so far and the cumulative trip length over the period yet to be simulated. Thus if the intended trip length $\bar{C}$ is known, and the simulation so far has yielded $\bar{C}(r)$, the trip length $\bar{\bar{C}}(r)$ to be attained during the rest of the simulation can be calculated by substituting $\bar{C}$ for $\bar{\bar{C}}$ in equation (7.12) and rearranging

$$\bar{\bar{C}}(r) = \frac{\bar{C} - \eta(r)\bar{C}(r)}{[1 - \eta(r)]} \quad . \tag{7.13}$$

However, the trip length $\bar{\bar{C}}(r)$ is the cumulative value required for the rest of the iterations whereas another possibility might be a trip length $\Delta\bar{\bar{C}}(r+1)$ to be attained in the following iteration or time period.

To find a value of $\Delta\bar{\bar{C}}(r+1)$ which would give a cumulative trip length of $\bar{C}$ at the end of the following time period, it is necessary to separate the cumulative trip length in equation (7.9) into two components. Then

$$\bar{C} = \bar{C}(r+1) = \frac{\eta(r)\bar{C}(r) + \Delta\bar{\eta}(r+1)\Delta\bar{\bar{C}}(r+1)}{\eta(r) + \Delta\bar{\eta}(r+1)} \quad , \tag{7.14}$$

where $\Delta\bar{\eta}(r+1)$ is the amount of activity to be allocated in time period r+1. $\Delta\bar{\eta}(r+1)$ depends upon future movers which have not yet been determined and thus it is necessary to approximate this value. A best approximation is given by

$$\Delta\bar{\eta}(r+1) = \sum_i [\Delta S_i^m(r) + \Delta S_i^s(r)]/E,$$

and clearly the goodness of the approximation depends upon the set of $\alpha_i(r+1)$ values (compare equation (7.3)).

It is now possible to compute the target $\Delta\bar{\bar{C}}(r+1)$ required for r+1 from

$$\Delta\bar{\bar{C}}(r+1) = \frac{[\eta(r) + \Delta\bar{\eta}(r+1)]\bar{C} - \eta(r)\bar{C}(r)}{\Delta\bar{\eta}(r+1)} \quad . \tag{7.15}$$

It is clear from equation (7.15) that if $\Delta\bar{\bar{C}}(r+1)$ is actually met in r+1 then $\Delta\bar{\bar{C}}(\tau) = \bar{C}$, $\tau>r+1$. By analogy to equation (7.15), the equation for the required service demand trip length is

$$\Delta\bar{\bar{S}}(r+1) = \frac{[\eta(r) + \Delta\bar{\eta}(r+1)]\bar{S} - \eta(r)\bar{S}(r)}{\Delta\bar{\eta}(r+1)} \quad . \tag{7.16}$$

If a constant value for $\Delta\bar{\bar{C}}(\tau) = \bar{\bar{C}}$, $\tau>r+1$ is required, then it is clear that this would be the same as the cumulative trip length $\bar{\bar{C}}(r)$ in equation (7.13) and that for time period r+1, $\bar{\bar{C}}(r)<\Delta\bar{\bar{C}}(r+1)$ if $\bar{C}(r)<\bar{C}$ and $\bar{\bar{C}}(r)>\Delta\bar{\bar{C}}(r+1)$ if $\bar{C}(r)>\bar{C}$.

There are a number of issues which affect the choice of equation (7.15) rather than (7.13) as the estimate for the next time period trip target. Previous experience with numerical methods of searching for parameters of spatial interaction models consistent with some trip length statistic suggests that there is a tendency to find parameters which underestmiate the change towards the intended trip length (Batty, 1976). Moreover, because $\Delta\eta(r+1)$ is an approximation which is based on the maximum proportion of activity which can occur, equations (7.15) and (7.16) will be in error in the best way: that is, because $\Delta\eta(r+1)\leq\Delta\bar{\eta}(r+1)$, there is a potential opportunity for greater correction to the cumulative trip length than in the case where equation (7.13) is used. In fact, equation (7.15) is generally preferable in that it is based on the assumption that the trip length $\bar{\bar{C}}(r+1)$ will not be met whereas equation (7.13) assumes that $\Delta\bar{\bar{C}}(r) = \bar{\bar{C}}(r)$ will be met.

Finally, the trip lengths in this particular application tend to decrease in later time periods for constant parameter values (see Figure 7.1).

Thus a method based on the assumption that the trip length targets be recalculated in each time period is preferable. In the empirical work to be described below, a procedure for further overshooting the target has been incorporated on the assumption that the algorithm as applied will systematically underestimate the difference between the cumulative trip length calculated for r and the target required for r+1. The coefficient $\beta$ is designed to achieve this overshoot. Then

$$\beta = \{[\Delta\bar{\eta}(r+1)]^{\delta}\}/\{[1-\eta(r)]^{\delta}\},$$

where $\delta$ is a parameter which varies from 0 to $\infty$. When $\delta$ is large, $\beta \to 1$, and the effect is to reduce the overshoot. The final target trip lengths can now be written as

$$\left.\begin{aligned} \Delta^*\bar{C}(r+1) &= \Delta\bar{C}(r) + \beta^{-1}[\Delta\bar{C}(r+1) - \Delta\bar{C}(r)], \quad \text{and} \\ \Delta^*\bar{S}(r+1) &= \Delta\bar{S}(r) + \beta^{-1}[\Delta\bar{S}(r+1) - \Delta\bar{S}(r)]. \end{aligned}\right\} \quad (7.17)$$

A suitable value for $\delta$ and thus for $\beta$ is identified later in the empirical work, but it is now necessary to examine the way in which the bounds on these trip targets are fixed before the procedures used to move towards these trip targets are outlined.

THE COMPUTATION OF UPPER AND LOWER BOUNDS ON THE TARGETS.

The upper and lower bounds on the trip lengths to be achieved in subsequent time periods of the simulation are based on allocating the activity yet to be generated by the main model so that the trip length be maximised for the upper bound and minimised for the lower bounds. As Evans (1973) has so cogently demonstrated, this can be achieved by running the model given previously in equations (6.55) to (6.79) with $\mu_1(r)$, $\mu_2(r)$ set at $+\infty$ or $-\infty$,

or by formulating the model in its linear programming equivalent (Wilson and Senior, 1974). In this application, a special algorithm was devised based on the structure given in equation (6.55) to (6.79) but including linear programming type logic for the allocation instead of spatial interaction models. In each time period of the main model after the trip length targets have been established, the model of this section is run twice to allocate the remaining $[1-\eta(r)]$ activity so that the predicted cumulative trip lengths $\bar{\bar{C}}(r)$ and $\bar{\bar{S}}(r)$ are maximised and minimised.

However, formal linear programming models are not actually developed for such predictions as this would be extremely time-consuming. For example, in the application developed here for the Peterborough urban region, there are 65 origin and destination zones, thus there are $65^2 = 4223$ variables to be actively considered in the solution. The model is subject to 129 constraints and this transportation problem might have to be solved several times for both the work trip and service demand sectors for the time periods required to allocate the remaining $[1-\eta(r)]$ activity. Frankly, this is computationally quite impossible: in a typical 7 time period process, assuming that the bounds have to be evaluated at the end of the first and subsequent time periods excluding the last, there would be a total of (6 + 5 + 4 + 3 + 2) x 2 submodels x 2 types of bound = 80 linear programs to solve with the above dimensionality or less (if the dimensionality is reduced through the time periods as the constraints are met). Thus if the algorithm is to depend on such linear programming solutions, it looks like blowing up the very problem it was designed originally to solve.

However, the linear programming models would give true upper and lower bounds whereas what is required to make the method work is really only approximations to these. In fact, if the bounds are a bit tighter which any approximate

method would yield, this may be preferable as it would restrict the parameters to more reasonable values. Several possible approximations to the bounds could be developed. For example, Vogel's approximation (Hadley, 1962) could be used to give a conservative solution but more appropriate to this context is to use the same constraint procedure as used to solve the original model in association with the linear programming submodels.

In the original model, the gravity models used to allocate activity were treated as singly-constrained, thus solved directly and constraints were dealt with by assuming that constraint violations were turned into movers to be reallocated in the next time period. In effect, the same procedure can be used to approximate the linear programming solution: the linear program is assumed to be only subject to a set of origin constraints, and can thus be solved by inspection (note that ties are broken arbitrarily). Then the results of the allocation are assessed for destination constraint violations. If such violations occur, these are converted into movers to be reallocated in the next time period, and these zones which have met their constraints are removed from further consideration. Thus the dimensionality of the problem is successively reduced. In essence, the method requires that only enough origin activity be allocated in any one time period to meet but not exceed the destination constraints, and thus the problem can be solved by inspection. Moreover, the method of re-allocating surplus activity back to its origin is based on a ranking algorithm elaborated below and consistent with the idea of optimisation.

Rather than develop a completely new equation system based on equations (6.55) to (6.79), changes appropriate to the linear programming approximation to this system will be indicated. Assuming the bounds for $r$-1 are to be

established and given the usual input to time period r, the change in work trips $\Delta\hat{T}_{ij}(r)$ is first calculated from

$$\Delta\hat{T}_{ij}(r) = \begin{cases} [\Delta S_i^s(r-1) + \Delta S_i^m(r-1)] \text{ for } \underset{j}{\mathrm{OPT}}\{c_{ij}\} , \\ 0, \text{ otherwise.} \end{cases} \tag{7.18}$$

The optimisation over the travel costs to the destination zones from the origin zone i can be for a maximum or minimum in the quest for upper or lower bounds respectively. Then the amount of population, the constraint violation tests and the surplus activity is calculated as previously using equations (6.56) to (6.58).

However the surplus redistributing procedures in equations (6.59) to (6.61) are no longer suitable as a proportionate reallocation of movers back to their source is inconsistent with the notions of optimisation by linear programming. In order that the reallocation be consistent, it is necessary to establish the order of optimality for the surplus $\Delta_j^p(r)$. Thus it is necessary to rank the values of $c_{ij}$ from worst to best for any zone j according to the type of optimisation being pursued (maximisation or minimisation). The idea of the algorithm for reallocating the surplus back is to reallocate back from the worst towards the best degree of optimality until all the surplus is dealt with. A positive reallocation back to the origin only occurs of course if the forward allocation, that is $\Delta\hat{T}_{ij}(r)$, is positive.

These ideas are embodied in the following algorithm. First define $\hat{\hat{\Delta}}_i^p(r)$ as the amount of activity allocated back to the origin i and $\hat{\Delta}_i^p(r)$ as the cumulative amount of the reallocation. $\hat{\Delta}_i^p(r)= 0$ before the algorithm begins. Equation (6.62) is solved first and then in the order from j worst to best, each origin i is considered

$$\hat{\hat{\Delta}}{}^{p}_{i}(r) = \begin{cases} \Delta\hat{T}_{ij}(r) \text{ if } \Delta\hat{T}_{ij}(r) \leq \lambda^{-1}\Delta^{p}_{j}(r), \\ \lambda^{-1}\Delta^{p}_{j}(r) \text{ if } \Delta\hat{T}_{ij}(r) > \lambda^{-1}\Delta^{p}_{j}(r). \end{cases} \tag{7.19}$$

The population surplus, the cumulative origin surplus and the trips are now adjusted as follows

$$\Delta^{p}_{j}(r) = \Delta^{p}_{j}(r) - \lambda\hat{\hat{\Delta}}{}^{p}_{i}(r), \tag{7.20}$$

$$\hat{\Delta}^{p}_{i}(r) = \hat{\Delta}^{p}_{i}(r) + \hat{\hat{\Delta}}{}^{p}_{i}(r), \quad \text{and} \tag{7.21}$$

$$\Delta\hat{T}_{ij}(r) = \begin{cases} 0 \text{ if } \Delta^{p}_{j}(r) > 0, \\ \Delta\hat{T}_{ij}(r) - \hat{\hat{\Delta}}{}^{p}_{i}(r). \end{cases} \tag{7.22}$$

Equations (7.19) to (7.22) are iterated until all the population surplus has been reallocated back to its source, and then the ratio $\sigma_i(r)$ is calculated as

$$\sigma_i(r) = \theta_i(r)\hat{\Delta}^{p}_{i}(r), \tag{7.23}$$

where $\theta_i(r)$ is as defined in equation (6.60).

A similar procedure is used to compute employment. First the change in service demands $\hat{\Delta}S_{jk}(r)$ is calculated from

$$\Delta\hat{S}_{jk}(r) = \begin{cases} \Delta\hat{P}_{j}(r) \text{ for OPT}\{c_{jk}\}, \\ 0, \text{ otherwise}, \end{cases} \tag{7.24}$$

and then equations (6.64) to (6.66) are computed to determine the surplus $\Delta^{e}_{k}(r)$. Equation (6.70) is now solved. A worst to best ranking is made for positive flows from all origin zones j to a destination zone k, new surplus variables $\hat{\Delta}\,{}^{p}_{j}(r)$ and $\hat{\hat{\Delta}}{}^{p}_{j}(r)$ are defined and $\hat{\Delta}^{p}_{j}(r)$ is initialised to zero, Then in accord with the worst to best ranking, and in the order

of the origin zones j, the surplus is reallocated back using

$$\hat{\hat{\Delta}}^p_j(r) = \begin{cases} \Delta\hat{S}_{jk}(r) & \text{if} \quad \Delta\hat{S}_{jk}(r) \leq \gamma^{-1}\Delta^e_k(r), \\ \gamma^{-1}\Delta^e_k(r) & \text{if} \quad \Delta\hat{S}_{jk}(r) > \gamma^{-1}\Delta^e_k(r), \end{cases} \tag{7.25}$$

$$\Delta^e_k(r) = \Delta^e_k(r) - \gamma\hat{\hat{\Delta}}^p_j(r), \tag{7.26}$$

$$\hat{\Delta}^p_j(r) = \hat{\Delta}^p_j(r) + \hat{\hat{\Delta}}^p_j(r), \qquad \text{and} \tag{7.27}$$

$$\Delta\hat{S}_{jk}(r) = \begin{cases} 0 \text{ if } \Delta^e_k(r) > 0, \\ \Delta\hat{S}_{jk}(r) - \hat{\hat{\Delta}}^p_j(r). \end{cases} \tag{7.28}$$

At this point, the surplus $\hat{\hat{\Delta}}^p_j(r)$ has to be converted back once again into origin employment movers and this is done using equations (7.19) to (7.22) which determine another set of movers $\hat{\Delta}^p_i(r)$. The mover ratio $\alpha_i(r)$ is finally computed from

$$\alpha_i(r) = \sigma_i(r) + \theta_i(r)\ \hat{\Delta}^p_i(r), \tag{7.29}$$

and the rest of the sequence of original model equations is solved (from equation (6.74)). The whole system is iterated in this fashion until the convergence limit is met and at this point the cumulative trip lengths become the appropriate bounds.

From the maximisation problem, the upper bounds $\bar{C}^u(r)$ and $\bar{S}^u(r)$ are defined and from the minimisation problem, the lower bounds $\bar{C}^\ell(r)$ and $\bar{S}^\ell(r)$ are set. These bounds define the range of physically feasible mean trip lengths, all of which could be attained if necessary during the rest of the simulation. They are conservative bounds due to the nature of the approximation and due

to the fact that they are equal to the cumulative trip lengths rather than the trip lengths associated with the subsequent time period. Then the general simulation can continue if

$$\bar{C}^{\ell}(r) \leq \Delta\bar{\bar{C}}(r+1) \leq \bar{C}^{u}(r), \qquad \text{and if}$$
$$\bar{S}^{\ell}(r) \leq \Delta\bar{\bar{S}}(r+1) \leq \bar{S}^{u}(r).$$

If either of these conditions is violated, it is necessary to rework the allocation in the previous time period so that the model produces targets which are physically feasible, that is, within bounds. This necessitates a new set of parameters with values more appropriate to the previous targets than the previous set of values, and as such, it involves resolving the least-squares equations. If the targets are within bounds, the simulation continues but new parameters need to be assessed by solving the set of least-squares equations consistent with the new targets. It is to the solution of these equations that this discussion now turns, this being the final element in the algorithm before the experimental work is outlined.

MOVEMENT TOWARDS THE TARGETS: DIRECTIONS OF SEARCH.

Once the targets have been established, the least-squares criterion given in equation (7.1) can be set up, noting that $\Delta\bar{\bar{C}}(r)$ and $\Delta\bar{\bar{S}}(r)$ are substituted for $\bar{C}$ and $\bar{S}$ respectively. To minimise the sum of squares function with respect to the parameters, first define the generalised sum of squares function $Z(\underline{\mu},r)$ where $\underline{\mu}$ is a 1 x L row vector of parameters, $\mu_{\ell}$, $\ell=1,2,\ldots,L$. The sum of squares function is composed of a set of K elements, $f_k^2$, where $f_k$ is the difference between the target and the value sought on iteration r. Then

$$Z(\underline{\mu},r) = \sum_{k=1}^{K} f_k^2 , \tag{7.30}$$

and it is required to minimise $Z(\underline{\mu},r)$ with respect to $\underline{\mu}$.

For a minimum to exist, the first order conditions are

$$\frac{\partial Z(\underline{\mu},r)}{\partial \mu_\ell} = 0, \quad \ell=1,2,..,L, \quad \text{and}$$

the second order are

$$\frac{\partial^2 Z(\underline{\mu},r)}{\partial \mu_\ell \partial \mu_m} > 0, \ \ell, \ m=1,2,\ldots,L.$$

The second order conditions can be arranged in an L x L Hessian matrix which must be positive definite. From equation (7.30), the first order conditions give the normal equations which are stated as follows:

$$\frac{\partial Z(\underline{\mu},r)}{\partial \mu_\ell} = 2\sum_k f_k \frac{\partial f_k}{\partial \mu_\ell} = 0, \quad \ell=1,2,\ldots,L, \tag{7.31}$$

and although it is clear that the L equations in (7.31) are nonlinear, it it possible to approximate the function by linearising the set using a Taylor expansion. Expanding equation (7.31) to terms of the first order gives

$$\tfrac{1}{2} \frac{\partial Z(\underline{\mu},r)}{\partial \mu_\ell} = \sum_k \{f_k \frac{\partial f_k}{\partial \mu_\ell} + \sum_m \varepsilon_m [\frac{\partial f_k}{\partial \mu_\ell} \frac{\partial f_k}{\partial \mu_m} + f_k \frac{\partial^2 f_k}{\partial \mu_\ell \partial \mu_m}]\} \quad , \tag{7.32}$$

where $\varepsilon_m$ are the errors associated with the L parameters $\mu_m$, m=1,2,...,L. It is possible to disregard the third term on the right-hand-side of equation (7.32) for it is clearly of the order $O(\varepsilon_m^2)$ and for small $\varepsilon_m$ is insignificant. Thus equation (7.32) can be set equal to zero and rearranged using this further approximation.

First, define the Jacobian matrix $\underline{J}$ which is of order K x L with element

$J_{k\ell}$ given as

$$J_{k\ell} = \frac{\partial f_k}{\partial \mu_\ell} \quad .$$

Noting that $\underline{f}$ is a 1 x K row vector of function values, $\underline{\varepsilon}$ is a 1 x L row vector of error terms, equation (7.32) can be rewritten and set equal to zero using the above approximation

$$\underline{0} = \underline{J}'\underline{f}' + \underline{J}'\underline{J}\ \underline{\varepsilon}' \tag{7.33}$$

It is a simple matter to solve for $\underline{\varepsilon}$. Then

$$\underline{\varepsilon}' = -(\underline{J}'\underline{J})^{-1}\underline{J}'\underline{f}' \tag{7.34}$$

and the vector $\underline{\varepsilon}$ is added to the vector $\underline{\mu}$ which forms the basis of an iterative scheme used to find the minimum of equation (7.30). That is, $\underline{\mu}(n) = \mu(n-1) + \underline{\varepsilon}(n-1)$, and on each iteration, the matrix $\underline{J}$ is updated with respect to the new parameter $\underline{\mu}(n)$ and a new error vector $\underline{\varepsilon}(n)$ is computed until convergence. Note that the inverse $(\underline{J}'\underline{J})^{-1}$ will only exist if the number of criterion functions $f_k$ is greater than the number of parameters $\mu_\ell$, that is, $K \geq L$, for obvious reasons.

In the case where K=L, it is assumed that each parameter is associated with a single function, and thus the set of equations is completely consistent. Then equation (7.34) simplifies to

$$\underline{\varepsilon}' = -\underline{J}^{-1}\underline{f}', \tag{7.35}$$

and it is clear that this equation gives the error as a function of a linear approximation to the response surface. In fact, equation (7.35) has a similar structure to the Newton-Raphson equation, and if $\underline{f}$ were a vector of first derivatives, $\underline{J}$ would be a matrix of second derivatives - the Hessian matrix - and equation (7.35) would give the Newton-Raphson iteration. In previous applications, equation (7.35) has been termed

Newton-Raphson iteration but more strictly it is a version of Gauss' algorithm.

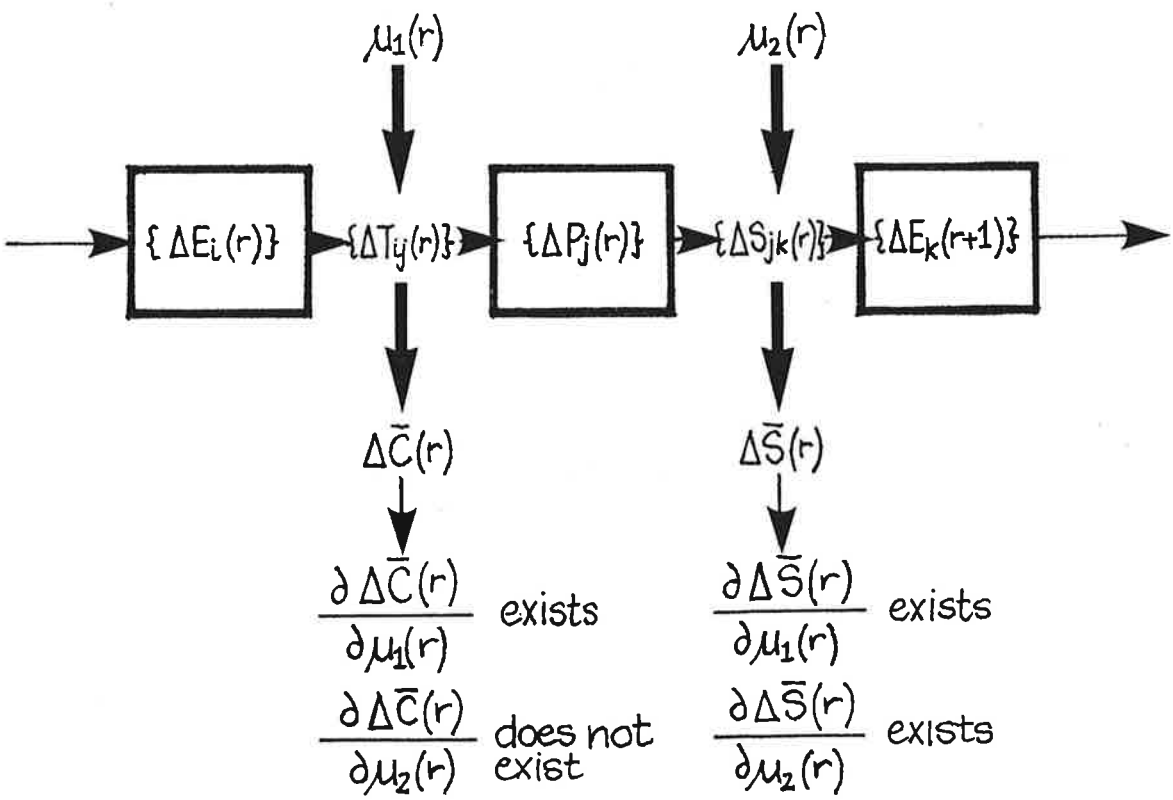
There are several advantages in using this least-squares approach to calibration. Clearly only a matrix of first order partial derivatives of the criterion function is required and this involves considerably less computation whether analytic or numerical approximation to the derivatives is used. Second, and perhaps more important is the fact that the method allows for the use of more functions than parameters: in short, it lets the system of equations be overdetermined and as such, it enables the introduction of weights on the significance or importance of each function $f_k$.

Finally, it can be shown that the Hessian matrix of partial derivatives associated with the function $Z(\underline{\mu},r)$ is positive definite, that this function is strictly convex and that the direction of search is 'downhill' (Kowalik and Osborne, 1968); this ensures that the method of solution will converge. There is, however, a further advantage which relates to the fact that the dependence between activities in the model is sequential rather than simultaneous in any one time period, and this considerably simplifies the computation and inversion of $\underline{J}$. To show this, it is now necessary to apply the method of solution to the two parameter model.

With K=2, L=2, the function $\Delta Z(r)$ in equation (7.1) is composed of two elements $f_1^2 = [\Delta\bar{C}(r) - \Delta\bar{\bar{C}}(r)]^2$ and $f_2^2 = [\Delta\bar{S}(r) - \Delta\bar{\bar{S}}(r)]^2$. The two parameters associated with these functions are respectively $\mu_1(r)$ and $\mu_2(r)$. The matrix $\underline{J}$ is as defined above for the two parameter, two function case, and thus equation (7.35) is used as a basis for solution. In the argument so far, it has been assumed that the functions $f_1$ and $f_2$ are interdependent

and that minimisation of the least-squares criterion would involve the simultaneous solution of two linearised normal equations. However, in any time period r, the function $f_1$ is computed first and although dependent on $\mu_1(r)$, is independent of $\mu_2(r)$. The function $f_2$ is dependent on both parameters as it is computed second.

The following diagram makes the structure of the model involving the computation of $f_k$ quite explicit:




It is clear from this structure of dependence that although $\Delta\bar{S}(r)$ depends on $\Delta\bar{C}(r)$, $\Delta\bar{C}(r)$ is independent of $\Delta\bar{S}(r)$ and thus the partial derivative

$$\frac{\partial f_1}{\partial \mu_2(r)} = 0.$$

Therefore, the Jacobian matrix $\underline{J}$ is lower diagonal and its inversion is simply a matter of solution by the method of forward substitution, rather than by any more involved algorithm such as Cramer's rule. For a 2 x 2 matrix, the actual decrease in computer time is small when compared to the total involved in the complete algorithm, but this idea points the way to the efficient calibration of several sector models which are linked in the kind of uni-directional sequence suggested by the above diagram. Indeed, this property is extensively exploited in the continuing elaboration of this algorithm and this is reported later in Chapters 8 and 9.

The final step in making the algorithm operational is to derive the 2 x 2 Jacobian matrix by taking partial derivatives of the trip length functions. Because the equation system is sequential and not simultaneous, it is possible to compute these derivatives analytically, rather than numerically and this is another advantage to the algorithm over its static equivalents (see Scheurwater, 1976). To evaluate these derivatives, it is worthwhile redefining the work trip and service demand interaction models associated with any time period r. From equation (7.6) the change in work trips is given as

$$\Delta T_{ij}(r) = E\Delta\eta(r)t_i(r)a_i(r)q_{jk}(r-1)\exp\{-\mu_1(r)c_{ij}\}, \quad i=k, \tag{7.36}$$

where $a_i(r)$ is the balancing factor which ensures that the model is origin-constrained. Then

$$a_i(r) = 1/\sum_j q_{jk}(r-1)\exp\{-\mu_1(r)c_{ij}\}, \quad i=k. \tag{7.37}$$

The service demand interaction is derived from equation (7.7) as

$$\Delta S_{jk}(r) = \lambda[\sum_i \Delta T_{ij}(r)]b_j(r)p_{ij}(r)\exp\{-\mu_2(r)c_{jk}\}, \quad i=k\ , \tag{7.38}$$

where $b_j(r)$ is given as

$$b_j(r) = 1/\sum_{i,k} p_{ij}(r)\exp\{-\mu_2(r)c_{jk}\}, \qquad i=k. \tag{7.39}$$

In evaluating the derivatives of $f_1$ and $f_2$ with respect to $\mu_1(r)$ and $\mu_2(r)$, the targets $\Delta\bar{C}(r)$ and $\Delta\bar{S}(r)$ are fixed and therefore independent of $\mu_1(r)$, $\mu_2(r)$. Thus noting the various independencies within the equation structure, the Jacobian matrix becomes

$$\underline{J} = \begin{bmatrix} \dfrac{\partial\Delta\bar{C}(r)}{\partial\mu_1(r)} & 0 \\ \\ \dfrac{\partial\Delta\bar{S}(r)}{\partial\mu_1(r)} & \dfrac{\partial\Delta\bar{S}(r)}{\partial\mu_2(r)} \end{bmatrix}, \tag{7.40}$$

Each of these partials can now be evaluated using equations (7.36) to (7.39) in the previously given trip length equations for $\Delta\bar{C}(r)$ and $\Delta\bar{S}(r)$

For the mean work trip length $\Delta\bar{C}(r)$, it is clear that the derivative is

$$\frac{\partial\Delta\bar{C}(r)}{\partial\mu_1(r)} = \sum_{ij}\sum \frac{\partial\Delta T_{ij}(r)}{\partial\mu_1(r)} c_{ij}/\sum_{ij}\sum\Delta T_{ij}(r), \tag{7.41}$$

which is made explicit by differentiating equation (7.36) with respect to $\mu_1(r)$ and substituting the result into equation (7.41). Then

$$\frac{\partial\Delta T_{ij}(r)}{\partial\mu_1(r)} = -\Delta T_{ij}(r)c_{ij} + \frac{\Delta T_{ij}(r)}{E\Delta\eta(r)t_i(r)}\sum_j\Delta T_{ij}(r)c_{ij}, \quad \text{and} \tag{7.42}$$

$$\frac{\partial\Delta\bar{C}(r)}{\partial\mu_1(r)} = \{-\sum_{ij}\sum\Delta T_{ij}(r)c_{ij}^2 + \sum_i\frac{(\sum_j\Delta T_{ij}(r)c_{ij})^2}{E\Delta\eta(r)t_i(r)}\}/\sum_{ij}\sum\Delta T_{ij}(r). \tag{7.43}$$

For the mean service demand trip length partially differentiated with

respect to the same parameter $\mu_1(r)$, equation (7.43) can be used again in equation (7.38). Then

$$\frac{\partial\Delta\bar{S}(r)}{\partial\mu_1(r)} = \sum_{jk}\sum \frac{\partial\Delta S_{jk}(r)}{\partial\mu_1(r)} \; c_{jk}/\sum_{jk}\sum\Delta S_{jk}(r),$$

and the appropriate differentials can be stated as

$$\frac{\partial\Delta S_{jk}(r)}{\partial\mu_1(r)} = \lambda[\sum_i \frac{\partial\Delta T_{ij}(r)}{\partial\mu_1(r)}]b_j(r)p_{ij}(r)\exp\{-\mu_2(r)c_{jk}\}, \quad i=k, \qquad (7.44)$$

and,

$$\frac{\partial\Delta\bar{S}(r)}{\partial\mu_1(r)} = \{\lambda\sum_{jk}\sum [\sum_i \frac{\partial\Delta T_{ij}(r)}{\partial\mu_1(r)}]b_j(r)p_{ij}(r) \exp\{-\mu_2(r)c_{jk}\}c_{jk}\}/\sum_{jk}\sum\Delta S_{jk}(r), \quad i=k. \qquad (7.45)$$

Finally, using the same procedure as in equations (7.41) to (7.43) the partial differential of the service demand trip length with respect to its own parameter $\mu_2(r)$ is stated as

$$\frac{\partial\Delta\bar{S}(r)}{\partial\mu_2(r)} = \{- \sum_{jk}\sum\Delta S_{jk}(r)c_{jk}^2 + \sum_j \frac{(\sum_k \partial S_{jk}(r)c_{jk})^2}{\Delta P_j(r)}\}\sum_{jk}\sum\Delta S_{jk}(r). \qquad (7.46)$$

For higher order derivatives which might be required if the Newton-Raphson method were to be used, the recurrence formula for the derivatives of trip distribution models developed by Evans (1971) is appropriate.

At this point, all the elements of the algorithm for adaptive calibration have been outlined and it is now essential to demonstrate the use of the algorithm in relation to a practical application. There are several aspects of the use of the algorithm concerning the number of various types of iterations, the values of certain parameters and points for starting and finishing yet to be defined, and in the following section, the

experimental work involved in the fine-tuning of the method will be described.

## APPLICATIONS, EXPERIMENTS AND REFINEMENTS TO THE ALGORITHM.

The model and calibration algorithm have been tested using data for the Peterborough urban region. This region has been divided into 65 zones and the application is fairly typical in scale to many of the urban models developed during the last decade. In this sense, the algorithm would be appropriate to similar applications. A detailed description of one version of the Peterborough model and its data base is given in a related paper by the author (Batty, 1978) and is thus not described any further here. The various elements which characterise the algorithm and which must be set before the method is applied can be divided into two types: structural elements dealing with the presence or absence of some feature, and numerical elements dealing with the best values of certain coefficients affecting the adaptive nature of the calibration. Early on in the experimental work, it was decided to explore the structural elements first and having reached conclusions as to the efficacy of these, to then examine the effect of different coefficient values on the calibration.

Four different structural elements were identified: the presence or absence of any formal target overshoot as defined in equations (7.17), the use of equation (7.13) or (7.15) in assessing the target, the number of iterations of Gauss' algorithm in moving towards the targets (that is, in approaching the optimal value of the least-squares criterion), and the use of previous or new parameter values in starting to find the optimum value of the least-squares criterion. These four elements control the detailed form of the adaptive algorithm whereas the numerical elements

control the fine-tuning, that is, the conditions under which the algorithm best operates for this particular application. Two elements of the numerical structure of the algorithm have been defined: first, the value of a parameter controlling the acceleration or deceleration of the direction of search, and second, the value of the parameters used in starting the operation of the complete algorithm.

Before the tests of these elements are presented, one feature of the algorithm still remains to be described. If the trip length targets are found to be out of bounds, the model must be rerun for the appropriate time period using a different set of parameters in order that a set of targets be found which are within bounds. In fact, when the model is rerun, the previous trip length targets are reused and thus the parameters must be approximated in a different and better way for the model to yield different output. Here it is assumed that if an out-of-bounds situation occurs, the parameters are assessed by decelerating the direction of search by 50%; that is, by setting the new parameters as $\mu_1(r)= \mu_1(r)+\varepsilon_1(r)/2$, $\mu_2(r) + \varepsilon_2(r)/2$. The total number of reiterations of the same time period is 5 and if the parameters diverge outside the range set by $-179 \leq \mu_1(r)$, $\mu_2(r) \leq +179$, their values are reset to $1/\Delta\bar{\bar{C}}(r)$ and $1/\Delta\bar{\bar{S}}(r)$ respectively (if this has not already occurred).

The structural elements were explored first. The overshoot parameter $\delta$ was set equal to 1 and 5, thus implying a situation of overshoot ($\delta$=1) in terms of the equation preceding equation (7.17) and a situation of no overshoot ($\delta$=5). The two different equations for assessing the trip length targets were used and four different iterative solutions to equation (7.35) were tried based on 1, 2, 3, and 4 iterations of the equation. The fourth element involved starting each time-period solution

with the new or old (previous) values of the parameters; in the case of new values, $\mu_1(r) = 1/\Delta\bar{\bar{C}}(r)$ and $\mu_2(r) = 1/\Delta\bar{\bar{S}}(r)$. There are 32 different combinations of these elements and thus the model was run for each combination of these elements from two different sets of starting values: $\mu_1(1) = 1/\bar{C}$ and $\mu_2(1) = 1/\bar{S}$; and $\mu_1(1) = 0.1/\bar{C}$ and $\mu_2(1) = 0.1/\bar{S}$.

The results of these runs are presented in Table 7.1 where the relative efficiency of each run is given by the number of iterations taken to come within an acceptable limit of $\bar{C}$ and $\bar{S}$. In fact, the differences between these runs indicates the need to reiterate within the same time period due to being out of bounds, and the model can be run in a minimum of 6 time periods if the results are always within bounds. Table 7.1 is also organised so that each row reflects the number of iterations of Gauss' algorithm. Hence lower rows in the tables show results from runs which have much greater computer time. An estimate of the computer time taken on each run is included in brackets wherever the model has produced a solution.

From Table 7.1, it is eminently clear that the overshoot facility ($\delta$=1) is redundant in that better results are obtained when there is no overshoot ($\delta$=5). Furthermore, equation (7.15) is to be preferred to equation (7.13) in evaluating the trip length targets, and it appears that there is a slight advantage to starting each time period solution using new rather than the previous parameter values. The question of the number of iterations of Gauss' algorithm is more uncertain. On balance, it appears that 2 iterations are necessary to guarantee a solution in this context although from reasonable starting positions, only 1 iteration is required. However,

Table 7.1: Number of Iterations and Computer Time Associated with the Variation in Structural Elements of the Algorithm.

START FROM $\mu_1(1) = 1/\bar{C}$, $\mu_2(1) = 1/\bar{S}$

| Number of Iterations of Gauss' Algorithm | Targets based on equation (7.15) | | | | Targets based on equation (7.13) | | | |
|---|---|---|---|---|---|---|---|---|
| | New Start | | Old Start | | New Start | | Old Start | |
| | $\delta=1$ | $\delta=5$ | $\delta=1$ | $\delta=5$ | $\delta=1$ | $\delta=5$ | $\delta=1$ | $\delta=5$ |
| 1 | * | * | * | 6 (17) | * | 7 (18) | * | 7 (18) |
| 2 | * | 6 (20) | * | 6 (20) | * | 6 (20) | * | 6 (20) |
| 3 | * | 6 (26) | * | 6 (26) | * | 6 (26) | * | 6 (26) |
| 4 | * | 6 (32) | * | 6 (32) | * | 6 (32) | * | 6 (32) |

START FROM $\mu_1(1) = 0.1/\bar{C}$, $\mu_2(1) = 0.1/\bar{S}$

| Number of Iterations of Gauss' Algorithm | Targets based on equation (7.15) | | | | Targets based on equation (7.13) | | | |
|---|---|---|---|---|---|---|---|---|
| | New Start | | Old Start | | New Start | | Old Start | |
| | $\delta=1$ | $\delta=5$ | $\delta=1$ | $\delta=5$ | $\delta=1$ | $\delta=5$ | $\delta=1$ | $\delta=5$ |
| 1 | * | * | * | * | * | * | * | * |
| 2 | 10 (35) | * | 10 (35) | * | 10 (35) | 10 (35) | * | 10 (35) |
| 3 | 9 (40) | 8 (35) | 9 (40) | 8 (35) | 14 (57) | 9 (40) | * | 9 (40) |
| 4 | 8 (43) | 9 (46) | 8 (43) | 7 (40) | 8 (43) | 8 (43) | * | 8 (43) |

NOTES: Figures in brackets indicate computer time in seconds, others indicate number of iterations. * indicates that no solution is reached from these starting positions.

it appears that 2 iterations are necessary in theory and practice to establish the direction of the search and thus 2 iterations are preferred despite the greater computer time.

Having established certain structural characteristics of the algorithm, the values of the starting positions for the parameters and the speed at which the algorithm moved towards its targets were explored so that the limits of effectiveness of the adaptive structure of the algorithm could be assessed. The starting values for the parameters of the model were selected in the range from 0.01/80 to 1.75/80 and 8 values for $\mu_1(1)$ and $\mu_2(1)$ were selected in this range. Note that it is assumed in these experiments that $\bar{C} = \bar{S} = 80$. The acceleration parameter g accelerates or decelerates the direction of search by changing the effect of the error terms in the following way: $\mu_1(r+1) = \mu_1(r)+g\varepsilon_1(r)$ and $\mu_2(r+1) = \mu_2(r)+g\varepsilon_2(r)$. Eight values of this coefficient were also selected from the range $0.1 \leq g \leq 1.75$ and thus $8^2 = 64$ runs of the model were made in the quest to find the best combination of starting values and acceleration parameter.

The results of these runs are plotted on the grids shown in Figure 7.3(a) to (d) where two major response surfaces are shown: the surface based on the number of iterations required to reach a solution and that based on the closeness of the solution to the intended target values. From these graphs, it is clear that any starting value for the parameter between zero and the inverse of the intended target value gives meaningful results. But the values of $0.75/\bar{C}$ and $0.75/\bar{S}$ give best results when combined with an acceleration parameter which tends towards unity. Therefore, smaller rather than larger values of these parameters seem to give the best results and it is interesting to note that the best value of g appears to be 1

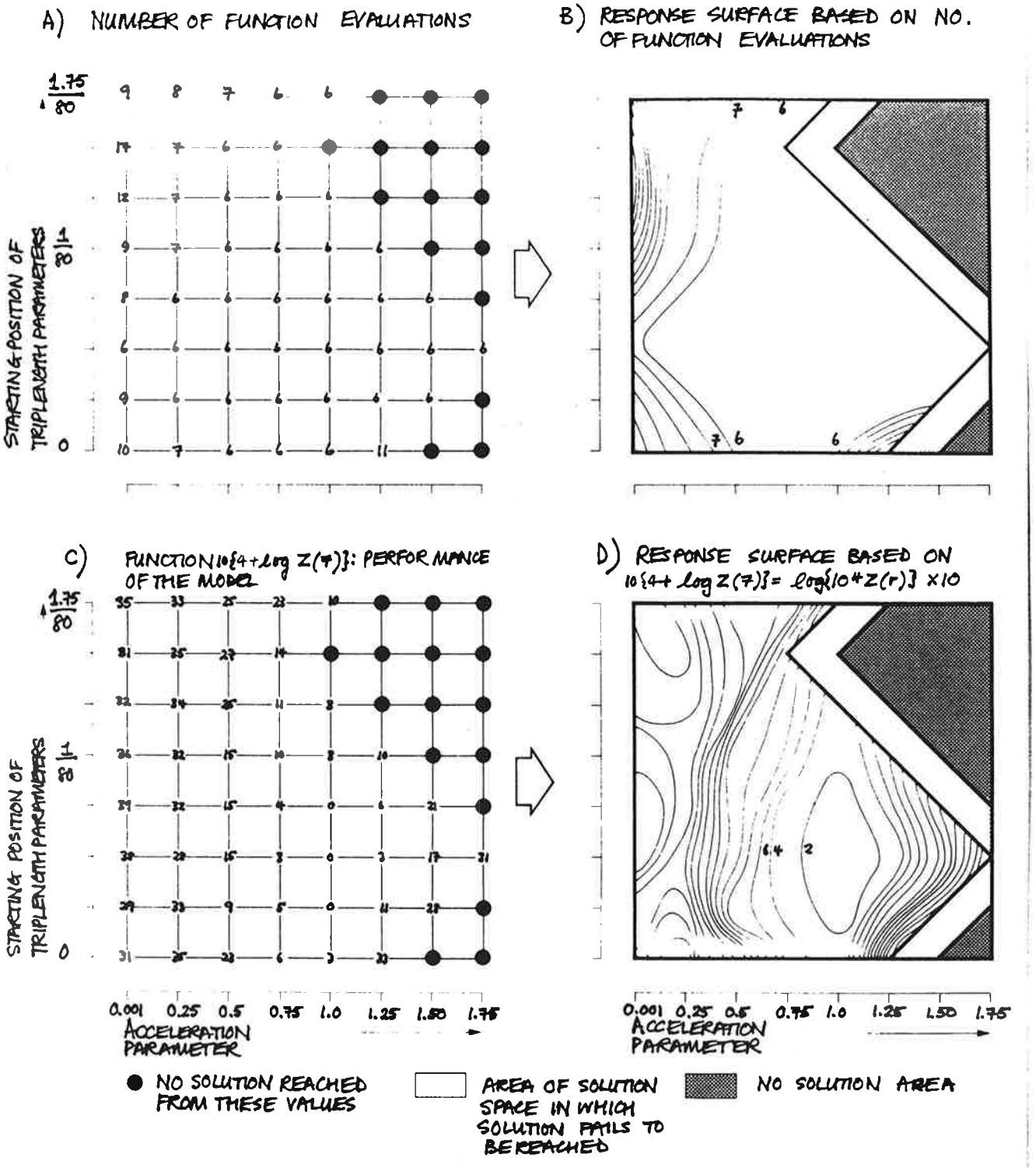



Figure 7.3: Response Surfaces Describing the Performance of the Calibration.

which implied that the acceleration parameter is redundant.

Finally, it is worthwhile showing a typical run of the model and the consequent solution procedure adopted by the algorithm so that readers may gauge its sensitivity to the target assessment, bounding and parameter solution procedures. To illustrate the algorithm, a typical starting point for the parameter values was adopted, and the model took some 10 iterations to reach the intended trip length targets. On the second iteration, the predicted targets went out-of-bounds three times before an acceptable within bounds target was found, and on the fourth iteration, the target went out-of-bounds once again. In fact, the simulation required 6 time periods before convergence but the out-of-bounds situation occurred 4 times making 10 iterations in total. Figure 7.4 illustrates the progress of the solution through its iterations in terms of the intended targets and the actual targets achieved. This figure also illustrates the four situations in which the out-of-bounds condition occurred and it is clear from this illustration that the solution procedure could be regarded as a kind of branch and bound procedure. A single path through the tree of potential solution paths is defined according to the bounding procedure and the degree to which Gauss' algorithm meets the required trip targets to each time period.

In Figure 7.4 it is difficult to represent the changing bounds on these graphs, but in Figure 7.5 a three-dimensional view of the solution procedure is presented in which the bounds are represented by the edges of the area contained in the solution space associated with each time period or reiteration of each time period. Here it is clear that on the third iteration, there is some oscillation in the violation of the bounds: the trip length targets first violate the upper, then the lower, then the upper

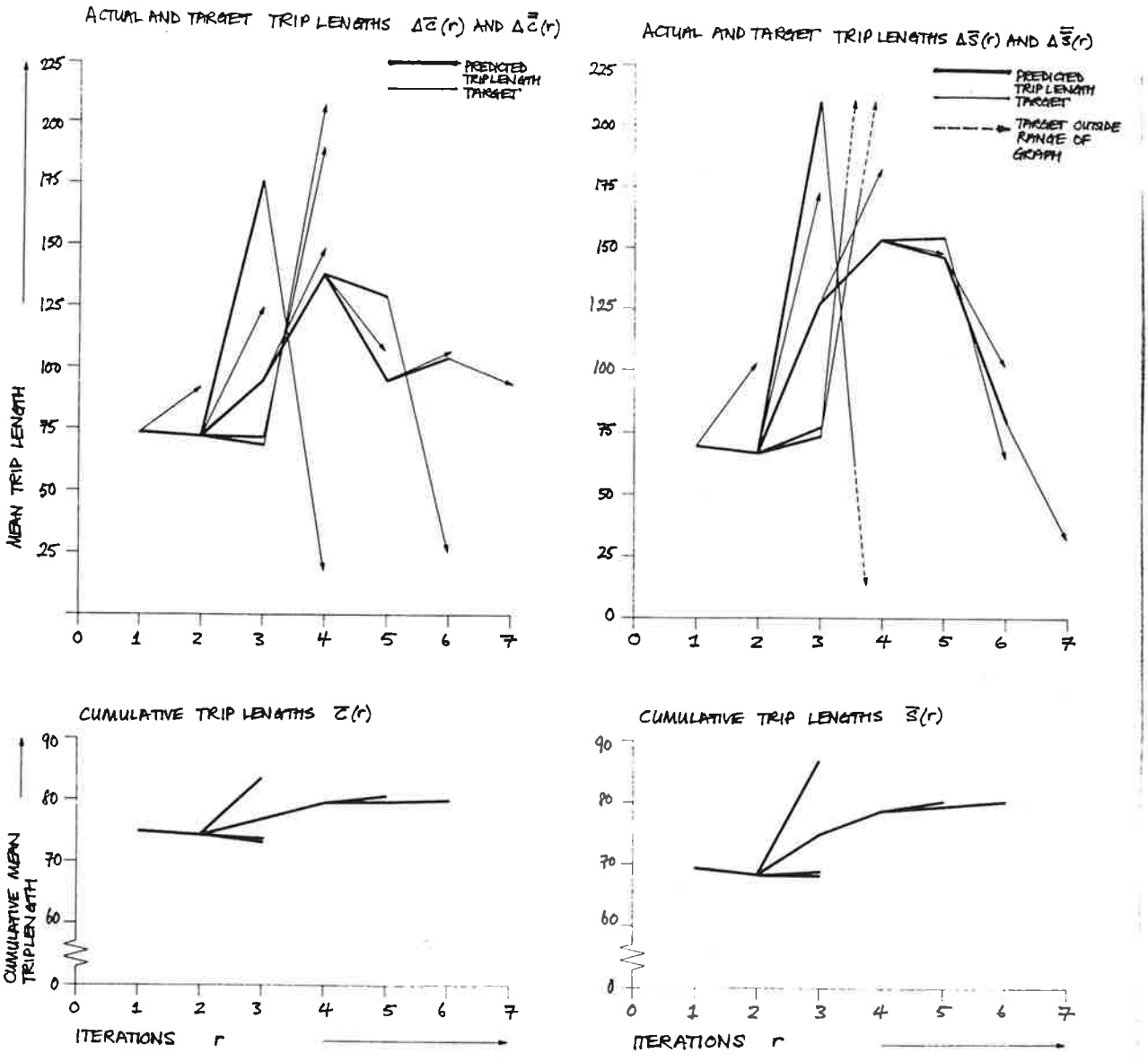



Figure 7.4: Actual and Target Trip Lengths Produced during a Typical Simulation.

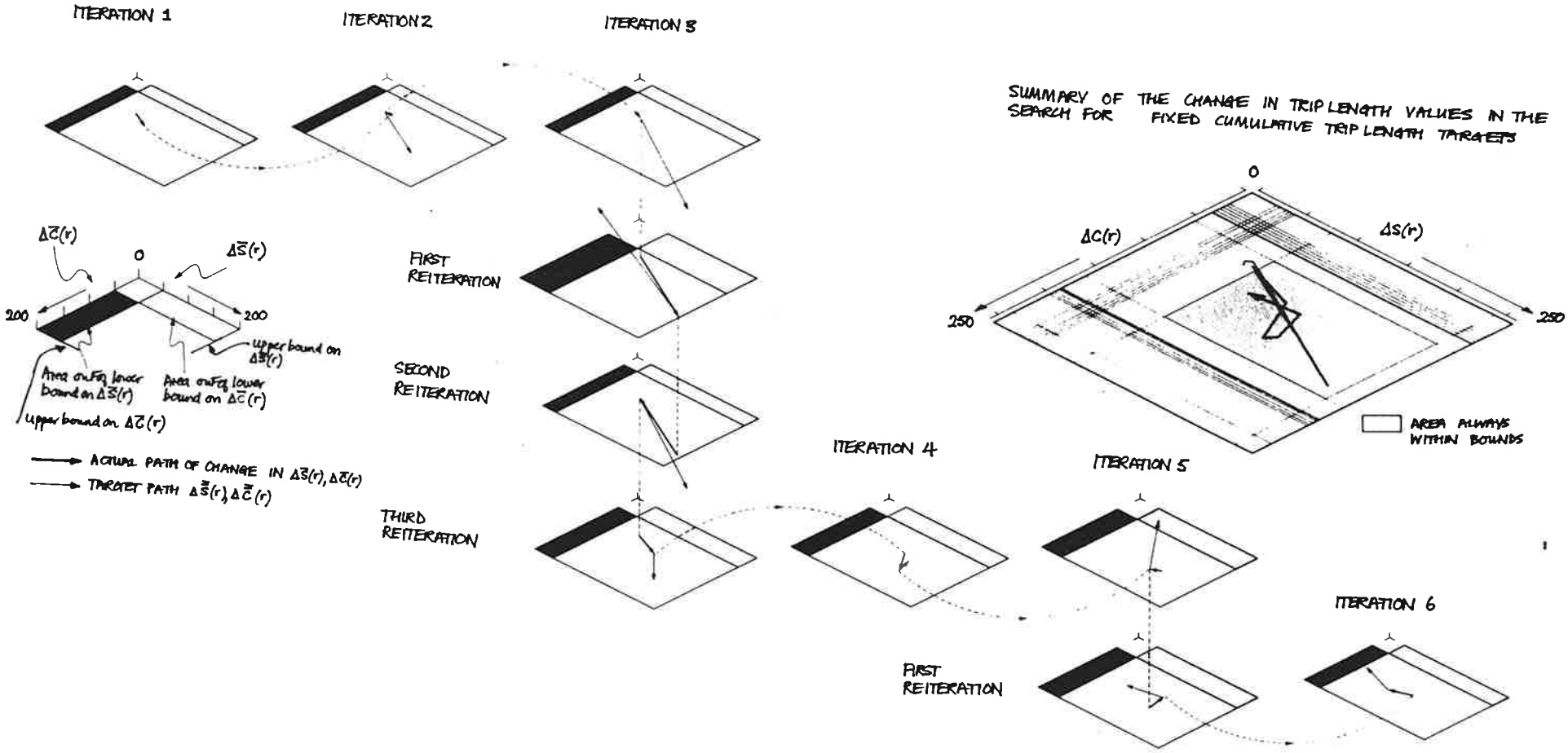



Figure 7.5: Upper and Lower Bounds on the Trip Lengths during a Typical Simulation.

bounds before an acceptable (within bounds) set of targets is achieved. Figure 7.5 contains a useful picture of the algorithm's adaptive properties in facing such a situation which, in this example, is dealt with quite successfully.

CONCLUSIONS.

No proposal for a new method of calibrating existing models with a pseudo-dynamic structure, would be complete without some statement of the efficiency of the method in comparison to existing alternatives. In fact, the alternatives are based on calibrating such a model in a static sense, by embedding the complete model within a wider iterative framework designed to find a set of parameters which yield the intended trip length targets. Such a method was developed in Chapter 5 and on a intuitive level, it would appear that any method which utilises the model's pseudo-dynamic structure would be preferable.

Yet there are complications to be considered: there is a considerable amount of additional computation required for the new algorithm and a count of FORTRAN assignment statements weighted to deal with different statement execution time shows that a typical time period iteration of the new algorithm takes 1.74 times as long as a typical time period simulation of the model of the previous chapter. To make the comparison explicit, a typical static calibration of the model of Chapter 5 involved 4 iterations of the Newton-Raphson method with an average of 6 iterations of the model's sequence. Thus 24 time units required for calibration must be compared to 6 x 1.74 = 10.44 time units for the fastest run reported in Table 7.1 (over twice as fast) and to 14 x 1.74 = 24.36 (just marginally slower) for the slowest run. These estimates are quite crude as they have not been

based on a strict experimental comparison but on a comparison of computer programs developed at different times for different problems. They do, however, go some way to showing that the algorithm reported here is preferable to established methods.

A more fundamental question emerges from this discussion for it is necessary to comment on the possibility of generalising this experience to other applications. Of course, this is a problem faced in the use of many methods whose ability to solve any problem cannot be definitively proved. Yet experience with the algorithm suggests that there can be a high level of confidence in applying it successfully elsewhere; for the problem to which it has been applied has many tricky characteristics which are not particularly favourable to the application of the algorithm (for a detailed discussion, see Batty, 1978). It is easier to generalise the particular ideas developed here rather than the complete algorithm for the notion of breaking up what at first sight, may appear to be a simultaneous structure, helps simplify the solution. Thus any model which has such a sequential structure at the micro-level and a simultaneous one at the macro can be simplified in this fashion. In particular, such a change in emphasis opens up new opportunities for using analytic rather than numerical derivatives and simplified methods of equation-solving.

There are parts of the algorithm which are quite cumbersome. The need to evaluate bounds is computationally time-consuming and potentially the weakest step in the chain of techniques necessary to successful operation of the method. Although the solution of the linear programs by inspection can be extremely fast due to successive updating of the optimal ranking of travel costs, and the use of the inverse ranking for the opposite bound problem, the method still absorbs some 60% of the additional time required

for one time period iteration of the model. There do, however, appear to be other possibilities for dynamic calibration which do not involve the calculation of bounds and these are being explored at present. For example, the complete mover model presented in the previous chapter, analogous to the model of Baxter and Williams (1975), does not involve any convergence of the model's process until a steady-state solution in which the constraints on location and calibration are satisfied. Thus the calibration can be achieved without the need for bounds which in the algorithm of this chapter are necessary because of the absolute convergence properties of the activity sequences. These possibilities will be explored in the next two chapters.

Although this chapter has largely concentrated on algorithms for calibration, the idea of the pseudo-dynamic model elaborated goes beyond notions of optimisation, and there are many substantive properties of these types of model yet to be elaborated. In future research, it will be necessary to examine the extent to which real dynamic processes can be approximated in the fashion shown here, for it is certain that the simple hypothesised sequences in these models can be made much more realistic. Moreover, an examination of the ways in which locational constraints are actually reached in dynamic urban systems might help in defining more realistic processes. Here it has always been assumed that a pseudo-dynamic model is defined only in the absence of some critical information inhibiting the specification of a fully dynamic model. Yet there are some instances where static models might be preferred due to the nature of the problem. For example, in cases where the focus is on marginal change, or where change is slow in any case - these may be situations where a pseudo-dynamic model will suffice, and such cases are also worthy of investigation in future research.

Finally, there appears to be great potential for examining the structure of the pseudo-dynamic process itself in formal terms. The possibility of systematically varying the life of such processes, the need to identify appropriate chains linking activities in sequential fashion in the manner implied in Chapter 2, the need to clarify, extend and refine the ideas behind spatial movers and stayers - these are some of the concepts which might be taken up in future research. There are many avenues to explore: existing models have and can be reinterpreted using these ideas but perhaps more important is the derivation of new forms of static model which are operationally feasible and simple to calibrate. Such models would be more coherently conceived and understood than existing static models, and their elaboration and application to specific situations would engender more relevant practice.

# CHAPTER 8.

## COMPLETE MOVER MODELS.

The interrelated philosophical and technical problems involved in building models of social systems are seldom more illuminated than in the study of dynamics. The general scientific method in which hypothesis and theory are in some way tested by experiment, either directly or indirectly on the system of interest, always seems to break down: relevant hypotheses which contain propositions about social behaviour are difficult to test empirically, and empirical analysis which does exist, is largely inductive and ambiguous in its support or rejection of specific theories. These difficulties are of course well-known. Data concerning change in social systems is often hard to obtain but more important are the intrinsic observational difficulties which occur when attempting to collect such information. A great deal of change in social systems will, by its very nature, go unrecorded forever, and this makes hypotheses concerning changes in social structures based on ideas about lags, leads, feedbacks and so on, exceedingly difficult, if not impossible to test. Moreover, what information is available is rarely sufficient to overcome the problems of ambivalence and equifinality which plague the interpretation of social phenomena. These problems make the prospect for dynamic modelling in the social sciences quite dismal, at least in the traditional sense, and as argued in earlier chapters, the preponderance of static theory and analysis is not surprising.

In urban modelling, the majority of models proposed have been static in conception but notwithstanding the difficulties already alluded to, there has been an inevitable tendency to speculate on appropriate dynamic forms. Such speculations have produced a diverse selection of approaches ranging from the somewhat wild, nontestable systems dynamics models originally proposed by Forrester (1969) to the much more careful conservative approaches implied in econometric urban models such as EMPIRIC (Irwin and Brand, 1965). This field like so many others in the social sciences has been torn apart by the dilemmas and paradoxes inherent in building dynamic models. On the one hand, dynamic models are theoretically essential due to their greater comprehensiveness and are thus intuitively more acceptable; on the other hand, static models are easier to build and test in practice. Commonsense suggests dynamics but feasibility implies statics. Most social scientists in the modern day appear to favour a realist position in which some form of empirical testability is necessary in the development of theory through modelling, and thus static models although severely limited, have become the order of the day.

The situation, however, is not as clear cut as these difficulties might imply. A whole range of urban models exist from fully-static to fully-dynamic and some progress is occurring in the design of models which are pseudo- or partially dynamic (Cordey-Hayes, 1972). This middle ground seems to be a promising areas for further work because it may be possible to eventually design relevant models which contain enough dynamics to be theoretically acceptable but are cast in a static framework which makes their structure testable in some sense. This idea exists in the work of Wilson (1981) where static spatial interaction models are embedded into dynamic processes, indeed are considered as the equilibrium outcomes of dynamics, and it is the central theme behind the ideas of this thesis.

Here static models within a dynamic framework and dynamic models within a static framework have been considered and in developing these ideas, a theory of dynamics sufficiently rich to enable temporal aggregation of various partially dynamic-static components was sketched. This was suggested in Chapter 3 in the light of the assumption that a comprehensive understanding of urban static structure involves a foray into dynamics. Firm support for this notion exists elsewhere: Samuelson (1948) in his seminal work *Foundations of Economic Analysis* also argues that "One interested only in fruitful statics must study dynamics".

The study of these ideas is clearly a much wider affair than the specific notions introduced in this or previous chapters. The discussion here will be orientated towards conventional static urban models but the essential argument of this chapter is once again to show how certain processes involved in such static models have a potential dynamic interpretation. This in itself is not new for there are many models for which such interpretations exist, but this argument will reinforce that of the last chapters in suggesting that these dynamic processes can be utilised to design better static models. In particular, static urban models require procedures for their calibration and for effecting their solution according to *a priori* constraints. Previously such procedures have been largely arbitrary, and have been computationally time-consuming and not particularly accurate. By exploiting the dynamic interpretation of static models, it is possible to design much more acceptable and faster procedures and this rather round-about logic leads to static models which are generally more relevant. In this sense then, a foray into dynamics very definitely leads to more 'fruitful statics' thus endorsing Samuelson's dictum.

This chapter will begin with a review of a conventional static urban

model with a well-known dynamic interpretation based on multiplier theory. Two ways of elaborating the dynamic structure are possible, the first based on a particular solution method for the model, the second based on aggregation from the pseudo-dynamic theory developed in Chapter 3. In both cases, the version of the static model resulting, enables concepts of spatial redistribution through the movement of existing activity (movers) to be quite cogently treated, and this suggests ways in which the dynamic process of the model can be controlled. The resulting model can also be interpreted as a cumulative multiplier model and in terms of previous chapters is a Baxter-Williams type model (see Baxter and Williams (1975) and Appendix 1). The notion of controlling the model's dynamic or rather pseudo-dynamic process (because time is only explicit, not essential) leads to methods for calibrating the model's spatial interaction functions and for effecting solutions which meet locational constraints. First, the constraints are handled using a biproportional procedure, and then calibration is treated using unconstrained optimisation: in essence, this involves matching the iterative structure of these methods to the dynamic (iterative) structure of the model in a manner similar to that in Chapter 6. An integrated algorithm based in biproportional and optimisation procedures is then developed. In Chapter 9, all these methods are tested on a small 10 zone problem taken from the LTS: London Traffic Survey (LCC, 1964) and then the final algorithm is tested on Central and West Berkshire data.

## CONVENTIONAL STATIC URBAN MODELS.

The highly aggregate model developed here and in previous chapters is characterised by two main variables: population and employment measured in terms of activities or the associated land uses containing these activities.

Employment is disaggregated into two main components - an independent component which is the prime input or driving force to the model, and a dependent component which like population is an output. In practice, quite detailed disaggregation of these two main variables could occur if required without altering the essential structural relations on which such a model is based. These relations are two-fold. First, there is a set of functional relationships which enable the level of the output to be generated from the given level of inputs; these are formalised by means of various multiplier relations. Second, there are spatial relationships in the form of spatial linkages or interaction which involve the way in which the output is spatially distributed from the input. The functional and spatial relationships are independent of one another and this has been a source of criticism in the past as has been the distinction between input and output employment which has been based on the traditional basic-non-basic split or some variation thereof.

Without loss of generality, it is now assumed that there are I spatial units or zones in the model and that population and employment are able to locate in each zone. Located variables are denoted by 1 x I row vectors and interaction is described in I x I matrices: vectors and matrices are shown in bold type (underlined here). For example $\underline{e}$ is a 1 x I vector of total employment and $\underline{p}$ a similar vector of population. Employment $\underline{e}$ is disaggregated into its independent component $\underline{b}$ and dependent component $\underline{s}$. In this chapter, it might be convenient to regard $\underline{b}$ as basic employment and $\underline{s}$ as service employment, although this is not essential to the theory of the model. This form of model is therefore of the economic base theoretic type and is thus related to the line of models originating from the model proposed by Lowry (1964).

In equilibrium, the two output variables population $\underline{p}$ and service employment $\underline{s}$ are simultaneously related through the model's functional and spatial relations. Then

$$\underline{p} = \underline{e}\,\underline{A}\ , \tag{8.1}$$

where $\underline{A}$ is an I x I matrix which translates employment into population through the functional relation of population dependence on employment and through the spatial relation organised about the journey to work. Service employment depends on population in the following way

$$\underline{s} = \underline{p}\,\underline{B}\ , \tag{8.2}$$

where the I x I matrix $\underline{B}$ fulfills the same role as $\underline{A}$ in that it translates population into services through a service demand function and its appropriate spatial dependence. The model is subject to the usual employment accounting equations

$$\underline{e} = \underline{s} + \underline{b}\ , \tag{8.3}$$

where $\underline{b}$ is the input employment defined here as basic employment.

The linear simultaneous form of the model is easy to demonstrate directly. Substituting equation (8.1) into (8.2) and the result into (8.3) leads to

$$\underline{e} = \underline{e}\,\underline{A}\,\underline{B} + \underline{b}\ , \tag{8.4}$$

from which the conventional reduced form is stated as

$$\underline{e}\,(\underline{I} - \underline{A}\,\underline{B}) = \underline{b}, \quad \text{or} \tag{8.5}$$

$$\underline{e} = \underline{b}(\underline{I} - \underline{A}\,\underline{B})^{-1}. \tag{8.6}$$

$\underline{I}$ is the I x I identity matrix. Equation (8.5) can be solved in a variety of ways: the well-known series expansion of the inverse $(\underline{I} - \underline{A}\,\underline{B})^{-1}$ has been used quite widely as a solution device in urban modelling (Batty, 1976) whereas direct matrix inversion appears to be more usual in macro-

economics (Dorfman, Samuelson and Solow,1958). Of interest here however is the class of methods referred to as matrix iterative techniques (Varga ,1962) which involve successive approximation to the stable value $\underline{e}$. Such methods are essentially iterative as their name suggests but they involve a potential dynamic interpretation which can be positively exploited in the solution of these sorts of model. Before deriving the appropriate method, it is worth stating how equation (8.5) can be solved using such techniques and this is presented below. Also note that in the following exposition, all the analysis is related to the equation for total employment $\underline{e}$ for in equilibrium, $\underline{p}$ and $\underline{s}$ directly follow from equations (8.1) and (8.2) once $\underline{e}$ is known.

It is possible to write equation (8.5) as

$$\underline{e}\ \underline{Z} = \underline{b}\ , \tag{8.7}$$

where $\underline{Z}$ is some I x I matrix relating total employment to the input $\underline{b}$. Then consider an appropriately dimensioned matrix split equation for $\underline{Z}$ defined as

$$\underline{Z} = \underline{U} - \underline{W}\ .$$

Using this split in equation (8.7) and rearranging to equilibrium form as in equation (8.4) leads to

$$\underline{e} = \underline{e}\ \underline{W}\ \underline{U}^{-1} + \underline{b}\ \underline{U}^{-1}\ . \tag{8.8}$$

The solution of equation (8.8) proceeds iteratively in that a new value for $\underline{e}$, say $\underline{e}$ (r+1), is computed from a previous value for $\underline{e}$, say $\underline{e}$(r), where the solution procedure is begun with some known or guessed value $\underline{e}$(o). Then

$$\underline{e}(r+1) = \underline{e}(r)\underline{W}\ \underline{U}^{-1} + \underline{b}\ \underline{U}^{-1}\ , \tag{8.9}$$

and it is clear that $\underline{e}$(r + 1) will only converge to $\underline{e}$ if $\underline{W}$ and $\underline{U}$ are

appropriately specified.

To examine the convergence, define the error on iteration r+1 as $\underline{\varepsilon}(r+1)$ which is computed as $\underline{e}(r+1)-\underline{e}$. A suitable recurrence equation for the error is

$$\underline{\varepsilon}(r+1) = \underline{\varepsilon}(r)\underline{W}\,\underline{U}^{-1}, \tag{8.10}$$

and in terms of the initial error $\varepsilon(o) = \underline{e}(o)-\underline{e}$, equation (8.10) becomes

$$\underline{\varepsilon}(r+1) = \underline{\varepsilon}(o)(\underline{W}\,\underline{U}^{-1})^{r+1}. \tag{8.11}$$

The error will only converge to the null vector if the matrix $(\underline{W}\,\underline{U}^{-1})^{r+1}$ converges to the null matrix $\underline{0}$. Formally, $\lim_{r\to\infty} \underline{\varepsilon}(r) \to \underline{0}$ if and only if $\lim_{r\to\infty}(\underline{W}\,\underline{U}^{-1})^{r} \to \underline{0}$, and the necessary and sufficient condition for this to occur is that the spectral radius (dominant eigenvalue) of $\underline{W}\,\underline{U}^{-1}$ be less than unity.

If this condition exists, then recursion on equation (8.9) leads to

$$\underline{e}(r+1) = \underline{e}(o)(\underline{W}\,\underline{U}^{-1})^{r+1} + \underline{b}\,\underline{U}^{-1} \sum_{v=o}^{r} (\underline{W}\,\underline{U}^{-1})^{v}, \tag{8.12}$$

and in the limit as $r \to \infty$, it is clear that

$$\underline{e} = \lim_{r\to\infty} \underline{e}(r) = \underline{b}\,\underline{U}^{-1} \sum_{v=o}^{\infty} (\underline{W}\,\underline{U}^{-1})^{v} . \tag{8.13}$$

The equilibrium solution using this type of analysis is also a series expansion which is analogous to the series expansion of the multiplier term $(\underline{I} - \underline{A}\,\underline{B})^{-1}$ in equation (8.6). The essence of this technique is, however, a matrix split which yields the appropriate condition on the matrix product $\underline{W}\,\underline{U}^{-1}$. Moreover, $\underline{U}^{-1}$ should have an easily invertible form such as a diagonal form and a good guess for $\underline{e}(o)$ will hasten the convergence. The so-called Jacobi split is based on such considerations: in this case, the split is already obvious and readers will have guessed that as $\underline{Z} = \underline{I} - \underline{A}\,\underline{B}$, $\underline{U} = \underline{I}$ and $\underline{W} = \underline{A}\,\underline{B}$. Substitution into equation (8.13) yields

$$\underline{e} = \underline{b} \sum_{v=o}^{\infty} (\underline{A}\ \underline{B})^{v} , \qquad (8.14)$$

which is the well-known matrix expansion of the multiplier $(\underline{I} - \underline{A}\ \underline{B})^{-1}$ due to Leontief (see Koehler, Whinston and Wright, 1975). Indeed if $\underline{e}(o)$ is chosen as $\underline{b}$ or as $\underline{o}$, the computational scheme implied by equation (8.12) is identical to the one used in many conventional applications of this urban model (Batty, 1976). This exposition of an alternative solution technique has added little as yet to the substantive interpretation of the model but the real interest in the idea relates to the way in which $\underline{e}$ is eventually approximated, and this involves the starting position $\underline{e}(o)$ and the solution path. It is these features which give the method its potential as a dynamic device to control the solution of models of this type.

DYNAMIC FORMS FOR STATIC MODELS.

Using the Jacobi split introduced above, it is clear that the matrix iterative equation (8.9) can be written as

$$\underline{e}(r+1) = \underline{e}(r)\underline{A}\ \underline{B} + \underline{b} \quad , \qquad (8.15)$$

which is a very different equation from the appropriate series expansion equation for $(\underline{I} - \underline{A}\ \underline{B})^{-1}$. In essence, equation (8.15) is regenerating and redistributing all the employment generated and distributed so far apart from basic employment which represents a fixed input distribution. But from what has been said already, it is clear that the matrix iterative approach will yield an equilibrium which is identical to that produced by other methods. To make the point, consider starting the solution of equation (8.15) with either $\underline{e}(o) = \underline{o}$ or $\underline{e}(o) = \underline{b}$. Then although the idea of regeneration and redistribution exists in a formal sense, the

characteristic series produced are no different from the more conventional iterative solution of equations (8.1) and (8.2) starting with $\underline{e} = \underline{b}$ .

For $\underline{e}(o) = \underline{o}$, equation (8.15) gives

$$\underline{e}(r+1) = \underline{b} \sum_{v=1}^{r+1} (\underline{A}\ \underline{B})^{v-1} ,$$

and for $\underline{e}(o) = \underline{b}$

$$\underline{e}(r+1) = \underline{b} \sum_{v=o}^{r+1} (\underline{A}\ \underline{B})^{v} .$$

The first series is lagged one period behind the second but whatever the starting position, both series will be identical to the conventional series expansion of the multiplier term when the limit is approached. Imagine, however, a process in which it is required to steer the solution towards some target which is specified in terms of the equilibrium state. Then this might be achieved by making the relationship matrices $\underline{A}$ and $\underline{B}$ time-dependent in some sense. For example, assume that $\underline{A}$ and $\underline{B}$ are defined at each iteration r as $\underline{A}(r)$ and $\underline{B}(r)$ and are so structured as to correct the state of the system given by $\underline{e}(r)$ in terms of the known target. In this way, feedback from the present state of the system to its future state is achieved and this is the essential basis of a system with adaptive behaviour.

In this model, equation (8.15) now becomes time-dependent in a distributional sense

$$\underline{e}(r+1) = \underline{e}(r)\underline{A}(r)\underline{B}(r) + \underline{b} , \qquad (8.16)$$

and in general, the equilibrium relations in equations (8.1) and (8.2) no longer hold. Recursion on equation (8.16) leads to a different form from equation (8.12)

$$\underline{e}(r+1) = \underline{e}(o) \prod_{t=o}^{r} \underline{A}(t)\underline{B}(t) + \underline{b} \sum_{v=1}^{r} \prod_{t=v}^{r} \underline{A}(t)\underline{B}(t) + \underline{b} \ , \qquad (8.17)$$

and assuming that $\underline{e}(o) = \underline{b}$ which henceforth will be the assumption of this paper unless otherwise stated, equation (8.17) becomes

$$\underline{e}(r+1) = \underline{b}\{\underline{I} + \sum_{v=o}^{r} \prod_{t=v}^{r} \underline{A}(t)\underline{B}(t)\} \ . \qquad (8.18)$$

This model is not just a series expansion of the multiplier with the constant matrices being replaced by time-dependent products, for the matrix products in equation (8.17) and (8.18) are taken backwards not forwards in time, this being characteristic of the regenerative and re-distributive nature of the process. A model in which there is no such redistribution but time-dependent change in the original relationship matrices has been suggested by Berechman (1976) and this would be specified as

$$\underline{e}(r+1) = \underline{b}\{\underline{I} + \sum_{v=o}^{r} \prod_{t=o}^{r} \underline{A}(t)\underline{B}(t)\} \quad , \qquad (8.19)$$

where the product is taken forwards in time, consistent with the way the series has been generated. In the case of models such as Berechman's, an equilibrium is likely to exist because the increments of activity generated by the series get smaller and smaller, eventually tending to zero. In this sense then the process would terminate. In the model of equations (8.16) to (8.18) however, the series is being progressively re-generated and redistributed and from Young's (1971) discussion of linear iterative methods, it is clear that the model is based on a non-stationary iterative method in which convergence of any kind can never be guaranteed in advance.

This same model has already been derived from rather different consider-ations which have much more substantive meaning. In Chapter 4, a class

of models was presented which were characterised by an initial generation and distribution of activity according to a series expansion of the form given in equation (8.14), and a series of regeneration and redistributions of this same activity in structurally similar series which start afresh in each successive time period. The amount of activity which is regenerated and redistributed is controlled by a mover ratio. Various types of model are derived by setting the ratio at different values, and the models are adapted for static situations by fixing the total number of stages characterising the initial series. These pseudo-dynamic models contain a particular model type which is derived when *all* the activity is regenerated and redistributed in successive time periods or iterations.

Under certain assumptions as to the form of the series expansion and assuming that the initial series is generated and distributed using the same time-dependent matrices which effect the regeneration and redistribution, this pseudo-dynamic model is equivalent to the model presented in equations (8.16) and (8.18). Furthermore, the model is similar to the one developed by Baxter and Williams (1975) in which they argue that the cumulative total of activity generated so far during a series expansion of the kind alluded to here, should be regenerated and redistributed, as the series is built up. In essence, the model of equation (8.19) can thus be interpreted as a cumulative, rather than incremental, economic base mechanism, although this only holds for models in which $\underline{e}(o) = \underline{b}$. In the sequel, this model will be referred to as the *complete mover model*; the more extensive and alternate derivation has already been given in Chapter 4 and Appendix 2 but readers are referred to Baxter and Williams (1975) and Berechman (1976) for a similar logic.

It is quite clear that the complete mover model is an example of a static model with a dynamic interpretation, yet the meaning of the dynamic process has not been explored in detail. As stated, the iterative or dynamic process is very largely a solution device in terms of matrix iterative analysis, and it might be considered that the temporal index of the process is more suited to computer time than historical time. Nevertheless, a more substantive interpretation does exist in terms of regeneration and redistribution and although such processes when operated in a static framework, might become solution devices to enable the system to meet some goal, target or constraint, the potential for a realistic approximation to historical processes does exist. Moreover, the very idea of dynamics in a static framework or vice versa is rather difficult to reconcile with pure statics or pure dynamics but such approximations appear necessary if the difficulties described in the introduction are to be minimised, and progress is to be made. At this point, it is necessary to return to an examination of the conditions under which the nonstationary complete mover model will become stationary for if this model is to be of more than academic interest, it must be capable of incorporating the equilibrium relationship which formed the starting point of this argument.

CONVERGENCE PROPERTIES OF THE COMPLETE MOVER MODEL.

All that can be said at present about the model given in equations (8.16) to (8.18) is that the process may continue indefinitely for it depends upon the sequence of matrices $\underline{A}(t)\underline{B}(t)$, $t = 0,1,2,\ldots, r$, which in turn depend upon considerations not yet stated. Thus an examination of convergence rests on an analysis of the matrices $\underline{A}(t)$ and $\underline{B}(t)$ which

contain the functional and spatial relationships on which the model is based. Earlier it was stated that these two relationships were separable, that is, that the functional relationship had no spatial implication and that the spatial relationship had no functional implication. In short, this assumption means that the functional relationships are independent of space and thus act as scalars controlling the total activity generated by the model. A particularly simple but widely used assumption is based on the idea that the functional relationship is temporally as well as spatially independent in the complete mover model, and it is the spatial relationship which pertains to time. Then the matrices $\underline{A}(t)$ and $\underline{B}(t)$ can now be partitioned as in earlier chapters

$$\left.\begin{aligned} \underline{A}(t) &= \underline{T}(t)\,\underline{\Lambda}\,, \qquad \text{and} \\ \underline{B}(t) &= \underline{\Gamma}\,\underline{S}(t)\,. \end{aligned}\right\} \quad (8.20)$$

$\underline{T}(t)$ is a row stochastic spatial probability matrix containing elements $t_{ij}(t)$ which measure the probability of an employee working in i and residing in j, $\underline{\Lambda}$ is a scalar diagonal matrix of inverse activity rates $\lambda$, $\underline{\Gamma}$ is a scalar diagonal matrix of population-service demand ratios $\gamma$, and $\underline{S}(t)$ is a row stochastic spatial probability matrix containing elements $s_{jk}(t)$ which measure the probability that a person residing in j will demand services in location k. Typically the interaction probabilities are modelled using submodels based on gravity or other model forms, and a widely used form of model has been the singly-constrained (singly-stochastic) gravity model popularised by Wilson (1970). Note however, that the framework is independent of the particular submodel used. The ratios $\lambda$ and $\gamma$ are usually calculated directly from exogenous data whereas the probabilities are calibrated numerically or estimated statistically as part of the model.

Using these assumptions in the cumulative economic base complete mover model given in equation (8.18) leads to a model form in which the effect of generation through the matrices $\underline{\Lambda}$ and $\underline{\Gamma}$ and distribution through $\underline{T}(t)$ and $\underline{S}(t)$ is much clearer. Then

$$\underline{e}(r+1) = \underline{b}\ \{\underline{I} + \sum_{v=0}^{r} (\underline{\Lambda}\ \underline{\Gamma})^{r-v+1} \prod_{t=v}^{r} \underline{T}(t)\underline{S}(t)\} \quad . \tag{8.21}$$

Because the two relationships relating to the functional generation of activities from the input and their consequent spatial distribution are so separate, it is possible to have two types of convergence in such a model. The first type of convergence which relates to generation is always assured by the way the process is defined. To demonstrate this point, it is necessary to aggregate equation (8.11) spatially and simply examine the generation characteristics of the process. Using the unit vector $\underline{1}'$ where the prime indicates transposition of the equivalent 1 x I row vector, equation (8.21) can be aggregated to

$$E(r+1) = \underline{e}(r+1)\underline{1}' = \underline{b}\{\underline{I} + \sum_{v=0}^{r} (\underline{\Lambda}\ \underline{\Gamma})^{r-v+1}\}\underline{1}' \quad ,$$

$$= \underline{b}\ \underline{1}'\ [1-(\lambda\gamma)^{r+2}][1-\lambda\gamma]^{-1} \quad ,$$

where $E(r+1)$ is the total employment generated by iteration $r+2$. For nontrivial urban systems, $\lambda\gamma < 1$ for this is the ratio of service to total employment, and thus as $r \to \infty$, $E(r+1) \to E$ which is a fixed level of total employment. Total population is also fixed from the equilibrium relationship and service employment is the difference between E and the total input which is fixed. Thus the model must always converge in terms of the absolute amount of activity generated.

It is of interest to examine the form of the model under specific

assumptions concerning the number of terms in the series required to approximate total employment E. Assume as earlier that this number is T + 1 units, thus implying that the length of time from the first term generated to the last is T units. Furthermore, make the assumption that each term in the series can only be regenerated T + 1 times, in other words, that there are T + 1 similar series used to regenerate and redistribute the activity, each of these series starting one stage after the last, and initially one stage after the first increment of activity in the original series has been generated. Clearly the last increment of activity will be regenerated and redistributed when r = 2T + 1, assuming that the process is begun with r = 0. Up to r = T, the appropriate equation is equation (8.19) or equation (8.21) if the assumptions on the matrices made above apply. From r = T + 1, however, the appropriate equation is

$$\underline{e}(r+1) = \underline{b}\{\underline{I} + \sum_{v=r+1-T}^{T+1} \prod_{t=v} \underline{A}(t)\underline{B}(t) + \sum_{w=o}^{r-T-1} \prod_{z=T+1}^{w+T+1} \underline{A}(z)\underline{B}(z)\}, \quad (8.22)$$

where the first sum and product term relates to activity which is still regenerating and redistributing itself and the second term involves the activity which is stable due to the termination of the appropriate mover sequence.

The first term is based on a backwards process and the second on a forward process and at the end of the modelling period (at 2T + 1), the final configuration of employment is calculated from

$$\underline{e}(2T+1) = \underline{b}\{\underline{I} + \sum_{v=T+1}^{T+1} \prod_{t=T+1}^{2T+1} \underline{A}(t)\underline{B}(t) + \sum_{w=o}^{T-1} \prod_{z=T+1}^{2T} \underline{A}(z)\underline{B}(z)\} \quad ,$$

$$= \underline{b}\{\underline{I} + \sum_{w=o}^{T} \prod_{t=T+1}^{2T+1} \underline{A}(t)\underline{B}(t) \} . \quad (8.23)$$

Equation (8.23) has the same form as equation (8.19) which is due to Berechman (1976); and thus a convergence has been reached which is the same as that involved in a model of conventional generation-distribution, with the final state not subject to the equilibrium relationships in equations (8.1) and (8.2). This form of model will not be taken any further here but it demonstrates the possibility that models of this type can easily be designed which converge to nonequilibrium states.

The second type of convergence which characterises the model is spatial or distributional. If at some time $t \geq T$, the spatial probability matrices $\underline{T}(t)$ and $\underline{S}(t)$ become independent of time and thus stable, that is, $\underline{T}(t)\underline{S}(t) = \underline{T}\ \underline{S}$, $t \geq R$, the model will converge in the following sense. Note that the process in which $\underline{T}(t)$ and $\underline{S}(t)$ actually do become stable need not be specified at this point but bear in mind the fact that the design of such a process and its optimal specification is one of the central tasks of this chapter. Then at time R from equation (8.16)

$$\underline{e}(R) = \underline{e}(R-1)\underline{A}(R-1)\underline{B}(R-1) + \underline{b}$$

$$= \underline{b}\{I + \sum_{v=o}^{R-1} \prod_{t=v}^{R-1} \underline{A}(t)\underline{B}(t)\}, \qquad (8.24)$$

where the process is started with $\underline{e}(o) = \underline{b}$. Then at and onwards from time R, the spatial probability matrices become stable and thus

$$\underline{A}(R)\underline{B}(R) = \underline{A}\ \underline{B} = \underline{T}\ \underline{\Lambda}\ \underline{\Gamma}\ \underline{S}\ .$$

An equation for n time units after R can be developed which demonstrates the convergence. Then

$$\underline{e}(R+n) = \underline{e}(R)(\underline{A}\ \underline{B})^n + \underline{b} \sum_{m=o}^{n-1} (\underline{A}\ \underline{B})^m\ , \qquad (8.25)$$

and it is quite clear that the nonstationary linear iterative process implied by equation (8.24) has been converted into the stationary linear iterative process of equation (8.25). In the limit as $n \to \infty$,

$$\underline{e} = \lim_{n \to \infty} \underline{e}(R+n) = \underline{b} \sum_{m=0}^{\infty} (\underline{A}\ \underline{B})^m$$
$$= \underline{b}(\underline{I} - \underline{A}\ \underline{B})^{-1} , \qquad (8.26)$$

and it is obvious that the model is consistent with the equilibrium relationships posed in equations (8.1) and (8.2). The idea of converting a nonstationary into stationary form is appealing: it has the potential for defining a process of feedback from the state of the system to its locational structure, which if stationary in itself, will enable the overall model to obtain an equilibrium. Defining such methods of feedback is also one of the central tasks of this chapter.

It is attractive to speculate that in theory, certain limits on generation and distribution are reached at different times. It is likely, for example, that an acceptable level of activity will be generated prior to an acceptable distribution of that activity being produced. However, it is in the nature of the notion of a limit that the process only approaches but never reaches such a limit and thus in practice, it is not possible to say that certain limits on generation are likely to be reached before distribution or vice versa. Moreover if the matrices $\underline{A}(r)$ and $\underline{B}(r)$ depend on $\underline{e}(r)$, then it is not possible to define stable $\underline{A}$ and $\underline{B}$ other than for $\underline{e}$, and thus the convergence of the generative and distributive processes cannot be separated. However, because these relationships are independent in a structural sense, convergence of one will not influence the other although the final

equilibrium needs to be specified in terms of their joint convergence. In models of this sort it is possible to attain exact convergence of the generation process using scalars which relate to the terms in the series expansion of the multiplier, and this would be desirable if the convergence were slow. However it is much more likely to be that the convergence is dominated by the methods used for reaching a stable distribution of activity and in this case, an exact solution cannot be obtained.

THE CONTROL OF PSEUDO-DYNAMIC PROCESSES.

The conventional static model of the type stated in equations (8.1) and (8.2) usually has to be manipulated so that its solution meets certain criteria relating to its spatial structure. This process of manipulation exists in two forms: first, there is the process of calibrating the model to achieve a certain performance in terms of its locational output and/or predicted interaction patterns. Calibration can be viewed in different ways - as a statistical process of achieving a best fit or as a numerical process of adjusting the model to reflect the characteristic orders of magnitudes of certain variables observed in the system. In this interpretation, the latter view is preferred and calibration is treated as a process of achieving correct dimensionality of the system. However, the essential point is that calibration usually involves a relatively small number of parameters in terms of the total output.

Second, there is the process or set of processes involved in making the model meet certain constraints on location which are in the nature

of feasibility constraints or policy constraints on the capacity of zones to receive activity. Possible methods for enabling the model vary from quite arbitrary schemes to much more elegant approaches which are consistent with the submodels. Usually in these cases, a much larger number of factors or parameters or arbitrary devices relative to the number of calibration parameters, have to be introduced to solve the model. For both calibration and constraints, these parameters or operations relate to the matrices $\underline{A}(t)$ and $\underline{B}(t)$, and in the example here, because $\underline{\Lambda}$ and $\underline{\Gamma}$ are specified exogenously, these parameters pertain to $\underline{T}(t)$ and $\underline{S}(t)$.

The general control problem can now be stated: given constraints on interaction and on location, the calibration and constraints procedures must determine the elements of the matrices $\underline{T}(t)$ and $\underline{S}(t)$ so that these constraints are met and an equilibrium solution attained. The essence of most methods for achieving this is to vary the elements in some trial and error fashion making corrections on the basis of the predicted state of the system. In the past, the usual practice has been to nest the complete equilibrium model, as given in equation (8.6) say, inside the constraint and calibration procedures, thus ensuring that the equilibrium is always met. A typical scheme for calibrating and constraining the model is worth describing for comparative purposes. The model as specified in equation (8.16) can always be solved, for $(\underline{I} - \underline{A}\,\underline{B})$ is, by definition nonsingular. It is possible to determine factors which relate to the matrices $\underline{T}(t)$ and $\underline{S}(t)$, and which reflect the operation of certain locational constraints: a well-known procedure for achieving this is biproportional factoring of the probability matrices $\underline{T}(t)$ and $\underline{S}(t)$ in which the factors are determined according to

the level of constraint violation. These biproportional or 'Furness Methods' are themselves iterative and thus the model must be nested within them (Bacharach, 1970; Evans, 1970). Finally the calibration parameters must be determined: methods of unconstrained optimisation have been used for this type of calibration quite successfully although these themselves are iterative, and thus the constraints procedure within which the model is embedded, must itself be embedded into the calibration scheme.

Computationally, such a nesting of iterative procedures is bound to cause problems for all but the smallest of applications and it is not surprising that a great deal of research has been devoted to speeding up the various iterative procedures. Typically, the matrix in equation (8.6) is approximated using three terms in the series expansion and an approximation for the rest (Batty, 1976). The biproportional constraints procedure is extremely slow, its slowness also being a function of the size of the problem, and in an urban modelling context, polynomial approximation to the factors has been suggested by the author and his colleagues (Batty, Bourke, Cormode and Anderson-Nicholls, 1974). In trip distribution modelling, Robillard and Stewart (1974) have attempted to use Newton's method in approximating the final form of these factors, but the problems they encountered led them to suggest that this method only be used to complement the Furness procedure. Finally, ways of speeding up the calibration based on accelerated Newton, and quasi-Newton methods have been explored (Batty, 1976). Some of these methods were developed in the last chapter.

It is worth presenting an example of the number of operations required

to calibrate this model using these methods. Assume that the basic operation is the calculation of one vector from another using a vector-matrix multiplication as in equations (8.1) or (8.2). Then there are 2 such operations for each term in the series expansion before the final approximation, making 6 in total. In the Berkshire model built by the author and his colleagues (Batty, Bourke, Cormode and Anderson-Nicholls, 1974) for testing some of these routines, 3 major iterations of the Furness procedure were used, supplemented by 3 applications of a doubly-constrained Furness procedure on each matrix after each iteration. These doubly-constrained models were calibrated in 5 iterations using the polynomial smoothing technique mentioned above. 2 iterations of the Newton-Raphson method were required for calibration in which the whole procedure was run 3 times to calculate numerically the partial derivatives of 2 parameters, and 1 final Newton-Raphson run was required after the procedure had been accelerated using unidirectional search. In total {[6 model matrix operations x 3 Furness iterations] + [3 doubly-constrained operations x 5 Furness-polynomial smoothed iterations]} x 2 Newton Raphson runs x 3 derivative calculation runs + 1 final Newton Raphson run = 141 matrix operations were required to calibrate the model. With a large model of 100 zones or more, a large amount of computer time is required and although this is substantially smaller than in earlier applications, it is still too high for standard use of a model such as this in a policy context. Moreover, the level of accuracy of the convergence is fairly coarse using this algorithm, and thus a better solution procedure is clearly required.

The way round the problems which arise from the conventional scheme

is already implicit in this argument. It was demonstrated in the previous chapter and it consists of utilising the dynamic structure of the model to achieve constraints on location and the calibration of parameter values. In essence, the idea is to match the dynamic (iterative) structure of the model with the iterative structures of the calibration and biproportional factoring procedures. In short, because an equilibrium implies that the matrices $\underline{T}(t)$ and $\underline{S}(t)$ be stable, and because the satisfaction of the constraints and interaction statistics implies that stable matrices exist, once stable matrices have been found, the model is in equilibrium and thus solved. Therefore, the new procedure can be seen as one in which the system is progressively steered towards its constraint levels and optimal parameters by altering the matrices $\underline{T}(t)$ and $\underline{S}(t)$ at each iteration of the model. Furthermore, the fact that the system is being built up at the same time if started with $\underline{e}(o) = \underline{b}$, leads to minimal constraint violations during the process.

In the last chapter, this type of logic was operated on the original series expansion without any regeneration or redistribution. Thus although the matrices were altered at each iteration of the model, the amount of activity generated at each iteration got smaller. In that model, at each iteration feasible upper and lower bounds on what could be achieved in the rest of the incremental process had to be established, and if the model went 'out-of-bounds', some backtracking was required. On average, in terms of the number of matrix operations required or their unit equivalent, some 125 operations were needed which compares favourably with 141 in the conventional algorithm. Yet it will be

shown here that by utilising the dynamic structure of the static model and by beginning the solution using a nonstationary method which ultimately becomes stationary, considerable progress over these levels can be made. Indeed, examples will be shown which demonstrate that the procedures in this chapter can be over 10 times as fast as conventional ones, and this could bring these types of model into more standard use.

There is another way in which models of this type can be solved more quickly using matrix iterative analysis. In such problems, anything which speeds up the solution is likely to be relevant, and if the equilibrium solution could be guessed accurately and the matrices fixed immediately, only one iteration would be required to establish equilibrium. A useful starting position would be the observed employment vector $\underline{\tilde{e}}$ so that $\underline{e}(o) = \underline{\tilde{e}}$ , and on the not-so-unreasonable assumption that predicted $\underline{e}$ is likely to be close to $\underline{\tilde{e}}$, such a starting position could be judicious. The closeness of $\underline{e}$ and $\underline{\tilde{e}}$ has been seen in conventional applications (Batty, 1976) and this represents a case of using all the information available to obtain the solution. Moreover, this idea could be useful in establishing a measure of closeness or fit between $\underline{e}$ and $\underline{\tilde{e}}$ in terms the $\underline{A}(t)$ and $\underline{B}(t)$ matrices rather than external statistics.

The use of this idea might pose problems in forecasting but at least it raises the notion that the present state can be explicitly used to reach the future state and as such, it opens the door to some interesting thoughts about complete mover fully-dynamic models and their calibration. All this is speculation, however, for there is a problem in the use of

$\underline{e}(o) = \underline{\tilde{e}}$ for constrained applications. It may be that certain locations become constrained which would not previously be so if $\underline{e}(o) = \underline{b}$ were used. In the solution of $\underline{e}$, once a constraint is violated, it has to be met and as this could also happen starting with $\underline{\tilde{e}}$, the test problem in the next chapter based on the London Traffic Survey data has been operated using basic employment as the starting vector. However, in the next chapter, some tests on the larger Berkshire model are reported which use $\underline{\tilde{e}}$ as the starting point. These tests although not definitive, do suggest that the problems of constraint referred to are marginal, and that these are out-weighted by increases in speed.

In the sequel, the submodels which determine the matrices $\underline{T}(t)$ and $\underline{S}(t)$ will be outlined first and this will complete the statement of the model as it is to be tested. In the following chapter the test results are then described: first the locational constraints procedure based on biproportional factoring is stated and tested quite separately from the calibration procedure. Some discussion of the calibration procedure is then presented relating to methods of unconstrained optimisation and their application to different sequences of the iterative process. The results of these runs are reported and this leads to the assembly of the integrated algorithm and its test. All the results are taken from the model as applied to the London Traffic Survey data which as discovered later was an excellent test problem. 10 zones are defined based on traffic sectors but the interaction between the sectors is considerable and a complete data base is available for the model in terms of basic and service employment, population, locational attractors, journey-to-work and service trip interaction

patterns and distance measures (LCC, 1964). The integrated algorithm is finally tested on a larger model of Central and West Berkshire based on some 63 zones, and this confirms the efficiency of the method.

SPECIFIC FORMS FOR THE SPATIAL INTERACTION SUBMODELS.

From the discussion so far, it is clear that the problem of constraining or controlling the model's solution relates not to the linear sequence which is determined by the aggregate properties of the recursive process but by the elements of the matrices which translate population into employment and vice versa. In particular, if the above assumption that $\underline{A}(t)\underline{B}(t) = \underline{T}(t)\underline{\Lambda}\ \underline{\Gamma}\ \underline{S}(t)$ is adopted, it is the matrices $\underline{T}(t)$ and $\underline{S}(t)$ which determine the form of the solution, and these must be examined accordingly. For the matrix $\underline{T}(t)$, a typical element $t_{ij}(t)$ is modelled using a singly-constrained gravity model which predicts the probability of an employee working in i and living in j at time or iteration t. Then

$$t_{ij}(t) = \frac{B_j(t)D_j \exp\{\mu_1(t)d_{ij}\}}{\sum_j B_j(t)D_j \exp\{-\mu_1(t)d_{ij}\}} \quad , \sum_i t_{ij}(t) = 1 \ , \qquad (8.27)$$

where $D_j$ is the locational attraction of residential zone j , $d_{ij}$ is some generalised measure of deterrence to travel (distance in this case) between i and j, and $\mu_1(t)$ is a parameter which controls the effect of the deterrence and thus the total amount of travel made in the system. $B_j(t)$ is a weight or factor which is applied to the locational attraction so that it is adjusted towards a level which meets a related constraint. $B_j(t)$ and $\mu_1(t)$ are the factors and parameter respectively, associated with the residential submodel and

which have to be determined by the constraints procedure and calibration during the iterative process.

An analogous model is used to predict the elements $s_{jk}(t)$ of $\underline{S}(t)$ and this too can be interpreted as a singly-constrained gravity model. $s_{jk}(t)$ is the conditional probability of the demand for services in location k, given that the demand originates at the residential location j, and this is modelled using

$$s_{jk}(t) = \frac{A_k(t)F_k \exp\{-\mu_2(t)d_{jk}\}}{\sum_k A_k(t)F_k \exp\{-\mu_2(t)d_{jk}\}} \quad , \quad \sum_k s_{jk}(t) = 1 \quad , \qquad (8.28)$$

where $F_k$ is the locational attraction of service centre k, $d_{jk}$ is a measure of generalised deterrence as used in the residential location model and $\mu_2(t)$ is an associated parameter controlling the total amount of service demand generated. $A_k(t)$ is a weight or factor which scales the locational attraction towards a level which implies that some constraint on location is satisfied, and thus $A_k(t)$ and $\mu_2(t)$ represent the factors and parameter to be calculated when the constraint and calibration procedures are applied to the model.

Using equations (8.27) and (8.28) it is possible to write the matrix iterative equation (8.16) in more explicit terms as

$$e_k(r+1) = \lambda\gamma\{\sum_i e_i(r)\sum_j t_{ij}(r)s_{jk}(r)\} + b_k \quad . \qquad (8.29)$$

Clearly the sequence of values $e_k(r)$, $e_k(r+1)$,... depends on the matrix elements $t_{ij}(r)$ and $s_{jk}(r)$ which in turn are only time-dependent in terms of the constraint factors and the parameters. For the model to reach an equilibrium, $t_{ij}(t)$ and $s_{jk}(t)$ must converge to $t_{ij}$ and $s_{jk}$

which are then independent of t. In equilibrium, equation (8.29) becomes

$$e_k = \lambda\gamma\{\sum_i e_i \sum_j t_{ij} s_{jk}\} + b_k \quad , \tag{8.30}$$

which is subject to the following constraints being met

$$\lambda \sum_j e_i t_{ij} \leq p_j^m \quad , \tag{8.31}$$

$$e_k - b_k \leq s_k^m \quad , \tag{8.32}$$

$$\sum_i e_i \sum_j t_{ij} d_{ij} / \sum_i e_i = \bar{C} \quad , \qquad \text{and} \tag{8.33}$$

$$\sum_i e_i \sum_j t_{ij} \sum_k s_{jk} d_{jk} / \sum_i e_i = \bar{S} \quad . \tag{8.34}$$

In the above equations (8.31) to (8.34), $p_j^m$ is the maximum level of population allowed in zone j, $s_k^m$ is the maximum level of service employment allowed in zone k, $\bar{C}$ is the observed mean amount of travel or mean work trip length in the system, and $\bar{S}$ is the mean amount of service demand: note that both these means are defined with respect to the frequency of interaction over distance.

Some explanation of the precise meaning given to these constraints is necessary before proceeding to outline methods for meeting them. Equations (8.31) and (8.32) are the population and service location constraints respectively and the factors $B_j(t)$ and $A_k(t)$ are associated with their satisfaction. The mean trip lengths in equations (8.33) and (8.34) relate to the amounts of travel generated by the two submodels, and this is controlled by the parameters $\mu_1(t)$ and $\mu_2(t)$ respectively. Clearly in terms of the submodels presented in equations

(8.27) and (8.28), for equilibrium to occur, these factors and parameters must also converge to stable values: that is, $A_k(t) \to A_k$, $B_j(t) \to B_j$ , $\mu_1(t) \to \mu_1$ and $\mu_2(t) \to \mu_2$ as the equilibrium is approached. The problem as set out in equations (8.30) to (8.34) is to be solved using the framework of matrix iterative analysis already introduced, but it is possible that there are other ways of solving the system. For example, if equation (8.30) were treated as a constraint equation, and an appropriate objective function specified, then the problem might be amenable to standard optimisation methods.

Indeed, the rather promising work along these lines done by Williams and Coelho (1977) was reported in Chapter 2. However there is little computational experience with such algorithms as yet and what does exist appears to suggest that nonlinear programming algorithms when applied to such problems can be extremely slow. Cesario (1973) demonstrates that the solution of a doubly-constrained gravity model by such algorithms can be in the order of 500 times slower than the more conventional interpolation methods for calculating the parameters embedded in a biproportional scheme. Some of Cesario's results were so bad that he did not consider it worthwhile to report them in his paper and although the problem with his algorithms may rest on the choice of a suitable objective function, his work does note bode well for the use of such techniques in practice, unless specifically adapted to the problem structure. And it is to these specific adaptations based on the problem structure that the methods introduced here pertain.

It is proposed to solve the system of equations given in (8.30) to (8.34) using the iterative scheme implied by equation (8.29), and to match this iterative procedure to the standard biproportional (Furness) procedure used to determine factors which are appropriate to the locational constraints; and to an unconstrained optimisation procedure based on Newton methods used to determine the parameters appropriate to the observed mean trip lengths. The obvious advantage of this scheme is that the Newton methods which involve matrix inversion are used for problems of small dimensionality whereas the biproportional methods which involve simple factoring are restricted to large problems. In this sense, the problem is partitioned into two related parts which are handled by two different techniques, and subproblem size is traded off against sophistication of the solution method. The use of these methods alongside the matrix iterative scheme enables corrections to be made to the total activity generated and distributed so far in terms of the constraint factors and the parameters so that the constraints in equations (8.31) to (8.34) are ultimately met.

No proof that these procedures will converge will be offered here. Proofs of convergence of the completely constrained biproportional problem exist (see Bacharach, 1970; Evans, 1970; Macgill, 1976) but a partially constrained problem of this kind is more uncertain. However, it is in the nature of the factoring procedure that the averaging inherent in the method will lead to convergence for a subset of equality constraints, and the procedure defined below will be specified accordingly. It is likely that a proof of such convergence could be offered for this problem but this is beyond the scope and emphasis taken in this paper. In the case of the calibration using

Newton's method or some such method, a proof of convergence depends upon convergence of the biproportional process, and upon convexity of the objective function. Although the experiments reported below reveal that the appropriate function is convex wherever it has been evaluated, no such proof can be offered although it is likely that convergence will always be obtained for typical problems.

In the next chapter, the biproportional solution of equations (8.31) and (8.32) will be first attempted assuming that $\mu_1(t)$ and $\mu_2(t)$ are given. Then the optimisation procedure in which $\mu_1(t)$ and $\mu_2(t)$ are chosen will be described assuming that the constraints on locations are non-applicable. Finally, both procedures will be tested together. In all the test runs reported, a limit of 30 iterations was fixed on the solution of equation (8.29) to conserve computer time. This as expected is a sufficient number to enable a comprehensive analysis to be developed, but to demonstrate more detailed convergence, the model has been run for 100 iterations. The final algorithm is ultimately tested on a more realistic problem which was also limited to 30 iterations.

CONSTRAINTS PROCEDURES BASED ON BIPROPORTIONAL FACTORING.

We will present the conventional biproportional procedure in this chapter before testing it in the next. First define two sets of zones, $Z^p$ and $Z^s$ which contain the set of constrained residential and service centre zones respectively. At the start of the model's operations when $r = 0$, $Z^p = Z^s = \Omega$, the empty set, and the factors $B_j(0) = 1, \forall_j$, $A_k(0) = 1, \forall_k$. At the end of each iteration, a test is made to see whether a residential and/or service centre zone

should be constrained and if this is the case, new factors are computed. Then for any residential zone j,

$$\text{if} \quad \lambda \sum_i e_i(r)t_{ij}(r) \geq p_j^m \quad , \quad \text{zone} \quad j \rightarrow Z^p \quad . \tag{8.35}$$

The operation described in equation (8.35) is performed for all zones j and on this basis, if a zone belongs to the constrained set $Z^p$, a new factor $B_j(r+1)$ is computed from

$$B_j(r+1) = B_j(r) \frac{p_j^m}{\lambda \sum_i e_i(r)t_{ij}(r)} \quad , \quad j \varepsilon Z^p \quad . \tag{8.36}$$

The same kind of operation is performed in relation to service centre zones: for any zone k ,

$$\text{if} \quad e_k(r+1) - b_k \geq s_k^m \quad , \quad \text{zone } k \rightarrow Z^s \quad , \tag{8.37}$$

and for all the zones belonging to $Z^s$ after equation (8.37) has been tested, new factors $A_k(r+1)$ are computed from

$$A_k(r+1) = A_k(r) \frac{s_k^m}{[e_k(r+1) - b_k]} \quad , \quad k \varepsilon Z^s \quad . \tag{8.38}$$

Note that zones which have not entered the constrained set are associated with factors which have remained unchanged from the first iteration, and are thus still equal to unity.

A number of points about this process need to be made. The term biproportional was first used generally by Bacharach (1970) in connection with the RAS method of adjusting an input-output table. To demonstrate that this method of factoring is equivalent to the row-column adjustment of a matrix such as an input-output table, it is necessary (as in this first set of test runs) to assume that the parameters are constant and

independent of time. Then for the residential location model $\mu_1(t) = \mu_1$, $\forall_t$, and it is now necessary to define the matrix to be biproportionally adjusted as the matrix of trips $T_{ij}(t)$. From equation (8.27), it is clear that on iteration r, the work trips $T_{ij}(r)$ are computed from

$$T_{ij}(r) = e_i(r)t_{ij}(r)$$

$$= a_i(r)e_i(r)B_j(r)D_j \exp\{-\mu_1 d_{ij}\}, \qquad (8.39)$$

where the factor $a_i(r)$ is defined as

$$a_i(r) = 1/\sum_j B_j(r)D_j \exp\{-\mu_1 d_{ij}\} \quad . \qquad (8.40)$$

Substituting for $B_j(r+1)$ from equation (8.36) into the equation for $T_{ij}(r+1)$ analogous to (8.39), it is easy to show that

$$T_{ij}(r+1) = \left\{ \frac{e_i(r+1)}{\sum_j T_{ij}(r) \frac{p_j^m}{\sum_i T_{ij}(t)}} \right\} \left\{ \frac{p_j^m}{\sum_i T_{ij}(r)} \right\} T_{ij}(r) \quad , \qquad (8.41)$$

where the two terms in the large brackets represent the appropriate proportional factors applied to the rows and columns of the trip matrix elements $T_{ij}(r)$ to determine $T_{ij}(r+1)$. Note that the process implied by equation (8.41) assumes that $T_{ij}(r)$ is first factored with respect to the columns and the resulting matrix is then factored with respect to the rows: the operations have been collapsed into a single equation although usually these are separated in more formal statementsof the method (see for example, Bacharach, 1970, or Evans, 1970).

Clearly what the procedure does is to scale up or down the appropriate locational attractors at the origin and destination towards the intended constraint values which are known. For example, in terms of the destination scaling operations implied by equations (8.36) and (8.38), the factors can be seen as weights which reduce or increase the attraction of the destinations to the location of activity. And such reductions or increases are required to move towards the intended constraint levels. The process is similar to the one suggested in economic equilibrium models of the Walrasian variety in which prices are reduced or increased according to whether excess demands are negative or positive (see for example, Scarf, 1973). The other point worth noting is that once a constraint has been violated, the above procedure ensures that constraint will eventually be met. This is achieved by testing to determine whether a zone should enter the constrained set if a violation occurs, and once in the set, the zone can never leave. The factoring is then determined on the basis of zones in the constrained set. It is an elementary point that constraint levels must be achieved once a violation occurs, otherwise convergence would never be possible.

The central question, of course, is whether or not the process will converge. As mentioned above, proofs are only available for the totally-constrained problem in which all rows and columns of the matrix are factored. A number of related proofs are now available (Macgill, 1975) and it does seem intuitively obvious that these could be extended to a partial set of row or column constraints. The process of averaging implicit in equation (8.41), for example, is

such that eventually the intended constraint levels are likely to be met, although convergence in such partially-constrained problems is likely to be much slower than in the totally-constrained. Difficulties might be encountered if the structure of values in the original matrix reflected some loop or cycle but in problems of this type, this is not possible. However, an extension of the various proofs already available to partially constrained problems would be useful as these methods are used quite extensively in urban modelling.

The speed of convergence is more of a problem because both theory and practice suggest that this is extremely slow. In methods of this sort which are essentially trial and error, slow convergence is to be expected but Robillard and Steward (1974) show that the slowness is also a function of the size of the problem. Clearly this is due to the fact that it takes more time in larger matrices for the row and column factoring to be felt throughout the system. But the speed of convergence which Robillard and Stewart demonstrate is horrific: they show that for a totally constrained problem, a conservative estimate for the percentage error in the total trip estimates to converge to within 5 digits precision from iteration to iteration, $10^7$ iterations would be required for a 100 zone problem. This is an upper bound for the number required and it would most certainly be less than this but this does give an indication of the problem to be faced. In practice, Cripps and Foot (1969) found the same problem in their Bedfordshire model and the constraint procedure was abandoned due to the slowness of its convergence.

To speed up convergence, Robillard and Steward try to anticipate the final values of the factors by expanding their mathematical form about existing values using Taylor's theorem, truncating at terms of the first order and thus solving the associated system of linear equations which is, in essence, the Newton-Raphson method. This, however, poses some real problems because of problem size and thus they resort to a matrix iterative method which they suggest should be complementary to the conventional biproportional procedure. Their work is interesting and although somewhat discouraging is probably worthy of further investigation in the context of this model in future research. There may be other algorithms of potential interest such as Scarf's (1973) techniques based on simplicial search: these too await further research.

Finally, it is worth noting that the way the matrix iterative solution procedure is structured, enables the constraints procedure to be applied as the system actually develops. Zones which reach their constraint limits first are constrained first and other locations are then affected by the fact that such constraints have been violated. Thus this is an improvement on previous methods in which the complete solution has been used as a basis for computing the two sets of factors. This also speeds up the procedure during the early iterations. In the test example, the constraints were set at 1.2 times the level of the observed values of population and service employment in the initial test results to be described below.

CONCLUSIONS.

We have developed the structure of the complete mover model in some detail in this chapter relating it initially to matrix iterative analysis and then to Baxter and Williams' (1975) work. The proposal which is to be tested in the next chapter is that locational constraints and calibration parameters are best optimised by 'completely moving' activity which is generated so far. Thus the essential difference between the algorithm to be developed in Chapter 9 and that in Chapter 7 is that total activity is to be regenerated and redistributed, rather than the increments of activity. In this chapter, we concluded with a statement of the biproportional factorising method. In the next, this method will be further developed and applied, then linked to the Newton-type methods presented in Chapter 7 to provide a general algorithm for model solution.

# CHAPTER 9.

# ALGORITHMS FOR EFFICIENT MODEL SOLUTION.

This chapter will integrate methods for enabling locational constraints to be met and parameter values to be optimised with the iterative structure of the complete mover model developed in Chapter 8. As such, this chapter should not be read in isolation from Chapter 8 for the equation systems and notation given here are those developed in Chapter 8. Locational constraints are handled using the method of biproportional factoring given in equations (8.35) to (8.41) rather than the surplus redistributing procedures which characterise the treatment of constraints in Chapters 4 to 7. However, the spatial interaction parameters will be optimised using Newton-type methods similar to those already applied in Chapter 7, and these methods will be restated in this chapter. The algorithms will be developed here using two test problems; first a simple 10 zone model built for data from the London Traffic Survey (LTS) pertaining to 1964 (LCC,1964). Second after refinements have been made to these algorithms from these tests, the general algorithm will be applied to a 63 zone model of the Central and West Berkshire region using data from 1971.

APPLICATIONS TO THE LTS PROBLEM.

A number of different variations on the biproportional scheme were tested in terms of the model, the first being the original method as implied by equations (8.35) to (8.38), the second being the introduction of a parameter which anticipates the level of constraint violation based on the stage reached in the generative process, and the third being based on fitting a polynomial function to the series of factors produced and subsequent extrapolation of their values. To demonstrate the relative efficiency of these methods, it is necessary to define a set of statistics which measure convergence. Two types of statistic were used here: a measure of the convergence of the factors themselves from iteration to iteration and a measure based on the stability of the final activity distributions from iteration to iteration. Both statistics were suitably normalised with respect to some known value to give them a percentage or ratio interpretation, and they are only applicable to the set of constrained zones. Then for the two sets of factors, $\{B_j(r)\}$ and $\{A_k(r)\}$, the two appropriate statistics are defined as

$$\theta^p(r) = \sum_{j \varepsilon Z^p} \frac{|B_j(r) - B_j(r-1)|}{B_j(r)} , \quad \text{and}$$

$$\theta^s(r) = \sum_{k \varepsilon Z^s} \frac{|A_k(r) - A_k(r-1)|}{A_k(r)} .$$

Note that as the equilibrium values of these factors are not known, the ratios only imply local, not global convergence.

To alleviate this problem, and to include a measure of the spatial

distributions predicted which are also affected by the amount of activity generated, the two statistics for the population and service employment distributions are defined as

$$\xi^{p}(r) = \sum_{j \in Z^{p}} \frac{|p_j(r) - p_j(r-1)|}{p_j^m} , \qquad \text{and}$$

$$\xi^{s}(r) = \sum_{k \in Z^{s}} \frac{|s_k(r) - s_k(r-1)|}{s_k^m} .$$

Acceptable convergence limits for these statistics appear to be in the order of $10^{-2}$. This means that from iteration to iteration, there is a 1% change in the total values of the factors and/or the activity distributions. If all the zones were constrained, this would imply that on average, the factor or activity were changing by 0.1% which can still be quite large absolutely for large populations or employments. For example, in a zone with a constraint limit of 200,000 persons on each iteration, a limit of $10^{-2}$would imply that on average this zone would still be changing by 200 persons on each iteration. In the following analysis, the results of all the methods developed will be ultimately shown by a summary table in which the maximum number of iterations for any of the four statistics presented above to come within limits $10^{-1}$, $10^{-2}$ and so on, are recorded. Such a table does not record the speed of convergence very accurately but the original method and the best method will be presented in more detail using convergence graphs.

As a baseline comparison, a doubly-constrained version of the model in which all population and service zones were constrained to their observed values was first run, and the acceptable limits of $10^{-2}$ were reached in 14 iterations. By the 28'th iteration, the limits of $10^{-5}$

had been reached. The original method was then run: it was not known in advance which constraints would be violated but ultimately (by iteration 30), 6 out 10 population zones and 9 out 10 service zones were constrained. The 10'th service zone is of course, constrained automatically due to the fact that there is a fixed total service employment to allocate; so in practice, all the service zones were constrained. The performance of the original method was quite poor in that the limit of $10^{-1}$ had only been reached by the 28'th iteration, and the slowness of the convergence is illustrated by a plot of the two sets of statistics on log-log graphs in Figure 9.1.

It is clear that the biproportional procedure although converging, 'wanders around' quite a bit due to the fact that it is only using information based on individual constraints and not their interrelations. In operating this model, activity is being generated at the same time as the correct levels of factors are being sought, and thus a run was attempted in which the factors were only computed after the 9'th iteration, a point at which 99% of the total activity associated with the test problem had been generated. By iteration 30, the $10^{-1}$ limits had not been met, thus it was clear that this method based on matrix iterative analysis, is immediately superior to conventional methods. Yet an examination of the factors themselves reveals that their form is remarkably regular. Figure 9.2 shows a plot of the two sets of factors and from this, it is clear that if the point at which zones become constrained could be anticipated in someway, then the procedure would certainly be speeded up.

A fairly obvious idea involving this notion of anticipation, is to

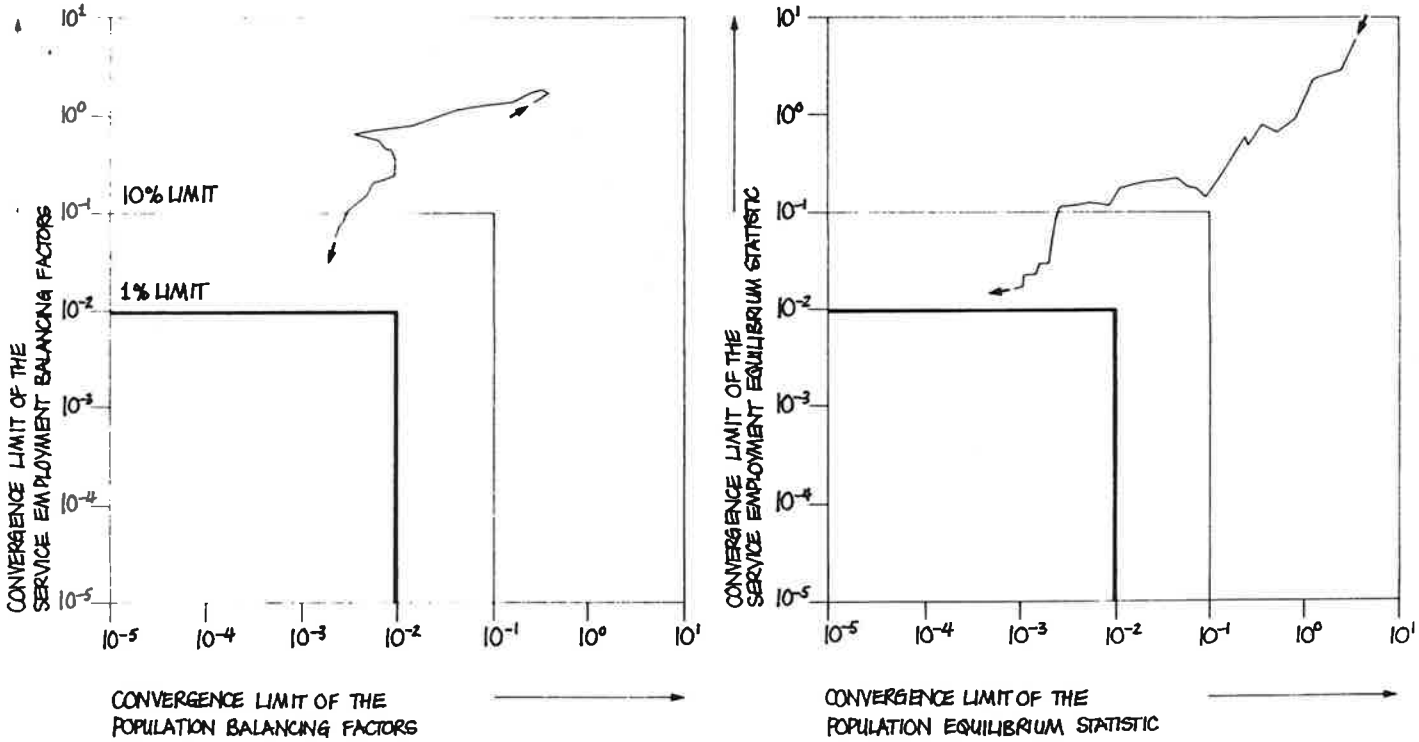



Figure 9.1: Convergence of the Original Biproportional Method.

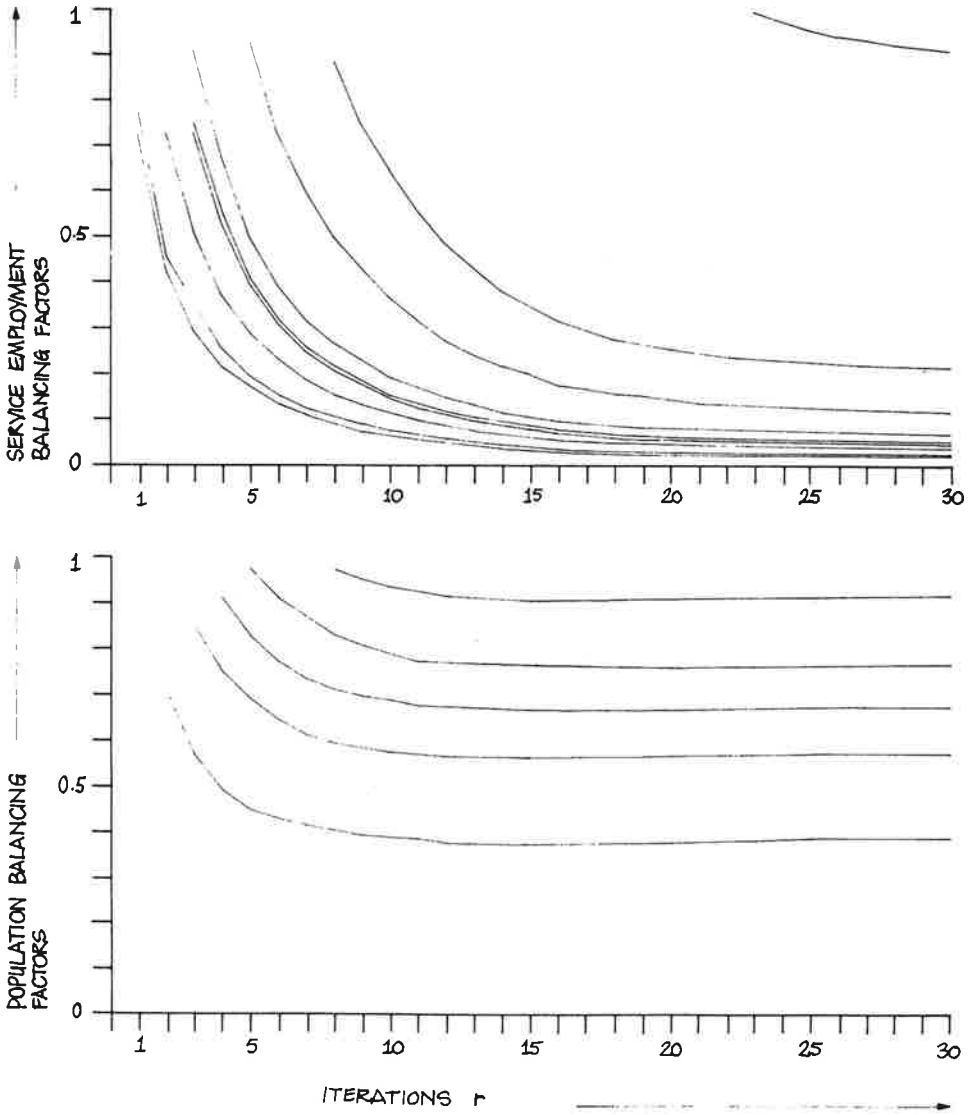



Figure 9.2: Convergence of the Biproportional Factors using the Original Method.

scale the amount of activity in proportion to its final total whenever a constraint is violated. For example, say that on the 3rd iteration, a constraint was violated but only 70% of the total activity in the model had been generated. On the assumption that if nothing were done about the constraint violation until most of the activity, say 99%, had been generated, it would be likely that this zone would receive more activity in proportion to the remaining 30% of activity to be generated. Thus by scaling the zone up after the 3rd iteration by the ratio of total activity to be generated to activity generated, an estimate of the eventual severity of this constraint could be computed. In general, on iteration r of the model, the percentage of activity generated so far is given by $[1 - (\lambda\gamma)^{r+1}]$ and the remaining activity is clearly $(\lambda\gamma)^{r+1}$. Thus the activities predicted so far must be scaled by $[1 - (\lambda\gamma)^{r+1}]^{-1}$ at each iteration to anticipate the amount of activity eventually locating in any zone. Then for the residential location model equations (8.36) and (8.38) are replaced by

$$B_j(r+1) = B_j(r)\frac{p_j^m[1 - (\lambda\gamma)^{r+1}]}{\lambda \sum_i e_i(r)t_{ij}(r)} \quad , \quad j\varepsilon Z^p, \quad \text{and} \qquad (9.1)$$

$$A_k(r+1) = A_k(r)\frac{s_k^m[1 - (\lambda\gamma)^{r+1}]}{e_k(r+1) - b_k} \quad , \quad j\varepsilon Z^s \; . \qquad (9.2)$$

Equations (9.1) and (9.2) can also be interpreted as a factoring scheme in which the constraints to be met are reduced to a level appropriate to the generative stage within which the model is, by the percentage of activity generated so far.

This method can only be applied when a constraint has actually been violated for if all activities were scaled, some constraints might be

violated unnecessarily. In essence the method is reasonably successful in that in the early iterations, it reduces the value of factors to a level closer to their ultimate values and thus speeds the process up slightly. Two versions of the method were tried: first the scaling procedure was applied throughout all 30 iterations of the model, and second, the scaling procedure was tested for sensitivity by applying it only after $[1 - (\lambda\gamma)^{r+1}] \leq 0.01$. It was not felt that this latter variation would give results any different from the original method but in fact, it makes a small difference due to the scale of activities involved. The method is operated over the whole sequence (30 iterations) and enables the limits of $10^{-2}$ to be reached on iteration 30, whereas although the second version does not, it is a very slight improvement on the original method as the later tabulation shows. Clearly, 30 iterations for a small problem is too slow, and therefore a third set of methods based on anticipating the final factor values was devised.

The regular pattern of change in the values of the constraint factors which are graphed in Figure 9.2 immediately suggest the possibility of fitting single and independent polynomial functions to a sequence of these values, and then extrapolating their equilibrium or limit values. A variety of functions could be developed based on different orders of polynomial applied at different stages of the iteration but as in all these methods, some balance between the advantages and disadvantages incurred by the technique must be sought. No exhaustive set of tests was begun in the search for an optimal polynomial fit but commonsense and the number of iterations which characterise the speed of the previous methods suggest that a quadratic form would be fairly relevant to test. For each of factors $\{B_j(t)\}$ and $\{A_k(t)\}$ different from unity, a quadratic of the form

$$B_j(t) = b_0 + b_1 t^{-1} + b_2 t^{-2} \quad ,$$

was fitted. Then as $t \to \infty$, $B_j(t) \to b_0$ which represents the equilibrium value. To fit such a function, three sequential values of $B_j(t)$ are required but in this instance, it was decided to fit the function based on average first-differences of these values, thus four values were required. In the case of $\{B_j(t)\}$, the three derived values were formed from $[B_j(r) + B_j(r+1)]/2$, $[B_j(r+1) + B_j(r+2)]/2$ and $[B_j(r+2) + B_j(r+3)]/2$. This particular version had already been tested by the author and his colleagues in the Area 8 modelling study (Batty, Bourke, Cormode and Anderson-Nicholls, 1974) and it was then found to be useful in smoothing localised 'kinks' in the sequence of factor values which tend to throw a non-smoothed method slightly off-course.

There are a number of difficulties with the application of any method such as this to the natural sequences implied by averaging methods of the biproportional type. The possibility of extreme limit values - negative or positive - exists: to counter this, $B_j(t)$ has been constrained to be between 0 and 1, the range within which it must lie. Negative values are set arbitrarily close to zero and values greater than 1 to 1. In no case so far have values greater than 1 been predicted, and the small number of cases where negative values have been predicted have posed no problems. The second problem is one of perturbation in processes such as this. Where polynomial functions are used regularly to anticipate equilibrium values, there is bound to be some disturbance to the natural factoring process which will show up in the model's predictions. The question of course is whether or not such perturbations significantly detract from the progress generated by the method.

In this case after four iterations of the model, if four values of the factors different from unity have been computed using the bi-proportional procedure, a polynominal is fitted, and a new value for use on the next model iteration is extrapolated. This new value however is associated with the last iteration, and thus it is only used to produce the value of the factor associated with the next iteration, and consequently, it is never used in the polynomial fitting four iterations later. Thus if the extrapolated value is off-course, it will be slightly modulated before use in future curve fitting and this reduces the degree of perturbation. Other possibilities are worthy of note: for example, it might be better to fit a polynomial of a higher order after a greater number of iterations, and to do this less frequently. It may be possible to overlap the fitting, thus continuously updating the limit value if a sufficiently good limit value can be initially established. A good deal of fairly basic work remains to be done in this area but on the basis of the results reported here, this appears to be a reasonably useful way to speed up the convergence of the biproportional process in problems of this kind.

Different ways of applying this quadratic fitting procedure were tested in the context of the model. The method was applied to factor values based on the anticipated generation described above in the second method, and on the original method. It was also applied to the total number of iterations and to the iterations only after the ninth when 99% of activity has been generated. Thus four varieties were tested in all, and all four come to within the $10^{-3}$ limits by the 28'th iteration, thus endorsing the relative and expected superiority of these methods. Table 9.1 records the relative convergence of all the methods introduced so far, and it is clear that the polynomial method in which the factors are anticipated

Table 9.1: Convergence of Various Biproportional Constraints Procedures in Terms of Maximum Number of Model Iterations Required.

| Type of Constraints Procedure | Doubly-Constrained | Partially-Constrained (Original Method) | Anticipated Generation (Second Method) | | Polynomial Method (Third Method) | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | No Anticipation | | Anticipation | |
| Convergence Limits of The Statistics | | | All Iterations | After Iteration 9 | All Iterations | After Iteration 9 | All Iterations | After Iteration 9 |
| $10^{-1}$ | 10 | 28 | 25 | 25 | 16 | 19 | 15 | 17 |
| $10^{-2}$ | 14 | - | 30 | - | 23 | 23 | 20 | 23 |
| $10^{-3}$ | 21 | - | - | - | 28 | 28 | 27 | 27 |
| $10^{-4}$ | 23 | - | - | - | - | - | - | - |
| $10^{-5}$ | 28 | - | - | - | - | - | - | - |

over the whole set of iterations is probably the fastest of all the partially-constrained methods. This is to be expected and a detailed illustration of the convergence of this method is presented in Figure 9.3 where the directness of the line of convergence must be contrasted to Figure 9.1 to gauge the performance. However, the degree of perturbation is also surprising, as is the fact that this perturbation does not seem to have much effect on the direction of the convergence. Another illustration of this point is given in Figure 9.4 where the values of the factors predicted by the biproportional and polynomial extrapolation are plotted. The discontinuities in the graphs are quite clear but the order of magnitude of the factors is established earlier than in the original method (compare with Figure 9.2) and by the 30'th iteration, these values are quite stable.

In general, the constraints on service centre location tend to dominate the overall convergence in this test problem. The statistics associated with convergence of the factors and distributions associated with services take longer to come within given limits than do the statistics relating to residential location, and this is due to the fact that ultimately all service zones are constrained in the solution. There is not much difference between the factor statistics and distribution statistics although the factor statistics appear to be marginally slower in their convergence. On balance, the polynomial methods are clearly best but in the integrated algorithms to be presented later, a variety of these methods will continue to be tested due to the fact that the integration of calibration and constraints procedures might lead to unforeseen advantages or disadvantages. At this point, the constraints procedures will be left and the emphasis will be on the calibration of the model without constraints in preparation for the discussion of an

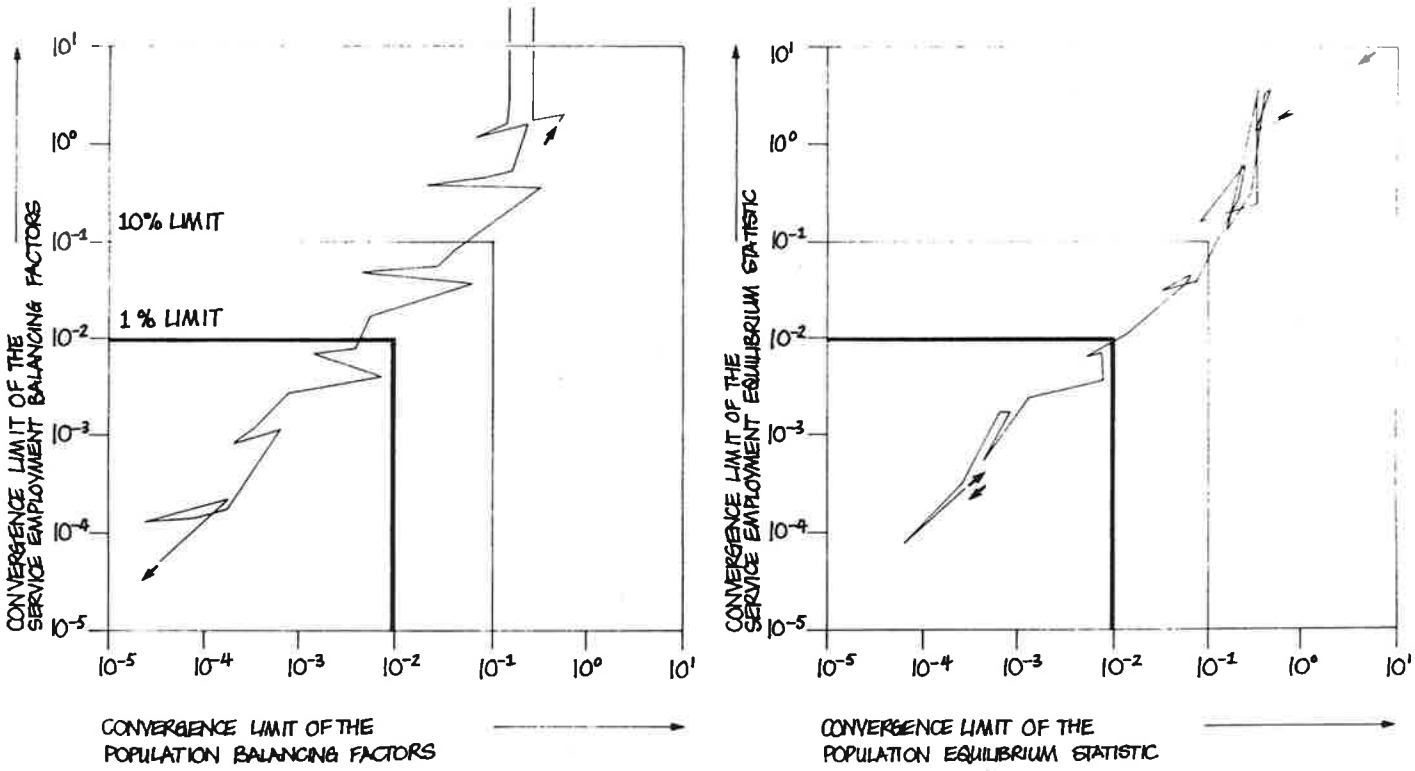



Figure 9.3: Convergence of the Polynomial-Anticipated Biproportional Method.

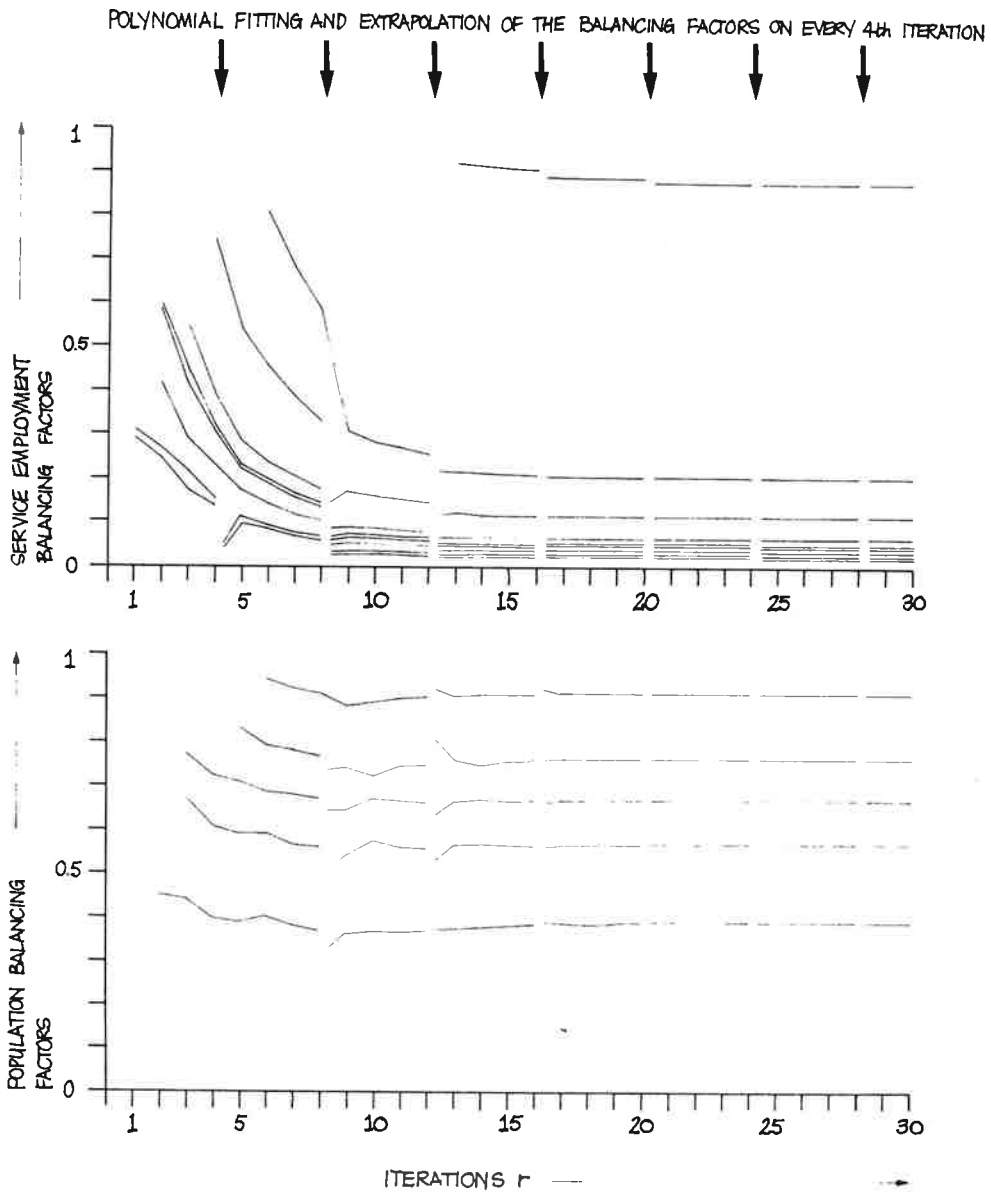



Figure 9.4: Convergence of the Biproportional Factors using the Polynomial-Anticipated Method.

integrated algorithm embodying the best elements of both procedures.

## CALIBRATION PROCEDURES BASED ON UNCONSTRAINED OPTIMISATION.

The main characteristic of the biproportional constraints procedures described above is the singular lack of any technique for measuring interdependence between each factor and between the sets of factors. Such interdependence clearly exists but the difficulty of accounting for it is primarily due to the size of the problem involved. In this test problem, the number of interdependent row and column factors is 36: 6 destination constraints in the residential and 10 in the service sector are ultimately required and 10 origin constraints in each sector are prespecified. Thus, to assess interdependence between 36 factors involves the systematic analysis of a 36 x 36 matrix of relations, and in most schemes in which some form of expansion or extrapolation of equilibrium values is required, a 36 x 36 matrix would have to be inverted or its inversion approximated. The work of Robillard and Stewart (1974) and the results of Cesario's (1973) research were sufficient to detract from exploring such methods at this stage. In contrast, the calibration problem involves a much smaller number of parameters, in this case 2, and usually, it is much more feasible to develop methods which emphasise their interdependence. In this section, the general basis of the techniques to be applied, will be presented and the two related techniques which are developed, will be structured in terms of the specific test problem.

In calibrating urban models, it is generally assumed that there are more statistics measuring the goodness of fit or numerical structure of the model than there are parameters whose values it is required to determine. Assume, therefore, that there are M statistics (or equations) and K

parameters (unknowns) where $M \geq K$. Construct the m'th statistic, $f_m$, from the deviation between some aspect of the model's prediction and its associated observation and square $f_m$ to get $f_m^2$, thus ensuring that this squared deviation statistic is always positive. Each statistic is dependent upon the K parameters, $\mu_k$, $k = 1, 2, \ldots, K$, as is the composite function $F(\underline{\mu})$ where $\underline{\mu}$ is the 1 x K row vector of parameters. This function is defined as

$$F(\underline{\mu}) = \sum_m f_m^2 \quad . \tag{9.3}$$

The calibration problem is thus one in which the sum of squares function in equation (9.3) is to be minimised by a judicious choice of parameters $\underline{\mu}$. The necessary conditions for a minimum can be stated by differentiating equation (9.3) with respect to $\mu_k$ and setting the result equal to zero. Then

$$\frac{\partial F(\underline{\mu})}{\partial \mu_k} = 2\sum_m f_m \frac{\partial f_m}{\partial \mu_k} = 0, \quad k=1,2,\ldots,K, \tag{9.4}$$

and the second-order conditions can be stated in a similar way.

However, it is possible to expand equation (9.4) about a known set of parameters values $\mu_k(r)$ using Taylor's theorem, and thus produce an approximation to equation (9.4) which provides the basis of a method for finding $\mu_k$. This of course is the basis of many such algorithms but the least-squares function has certain special characteristics which enable interesting and useful approximations to be derived. Approximating the expansion of equation (9.4) to the first order leads to

$$\frac{1}{2}\frac{\partial F(\underline{\mu})}{\partial \mu_k} \simeq \sum_m \left\{ f_m(r)\frac{\partial f_m(r)}{\partial \mu_k(r)} + \sum_\ell \left[ \frac{\partial f_m(r)}{\partial \mu_k(r)} \frac{\partial f_m(r)}{\partial \mu_k(r)} + f_m(r)\frac{\partial^2 f_m(r)}{\partial \mu_k(r)\partial \mu_\ell(r)} \right] \varepsilon_\ell(r) \right\}$$

$$\simeq 0,\ k = 1,2,\ldots,K. \tag{9.5}$$

If it it assumed that this approximation is correct, it is possible to write the system of equations implied by equation (9.4) in matrix terms and solve directly. Then

$$\tfrac{1}{2}\nabla\underline{F}' = \underline{J}_r'\underline{f}_r' + \{\underline{J}_r'\underline{J}_r + \underline{G}_r\}\ \underline{\varepsilon}_r' = 0\ , \tag{9.6}$$

where $\underline{J}_r$ is an M x K Jacobian matrix of first derivatives of the function $\partial f_m(r)/\partial \mu_k(r)$ , $\underline{f}_r$ is a 1 x M row vector of function values $\underline{f}_m(r)$, $\underline{G}_r$ is a K x K matrix of second-order terms $\sum_m f_m(r)\partial^2 f_m(r)/\partial \mu_k(r)\partial \mu_\ell(r)$, and $\underline{\varepsilon}_r$ is a 1 x K row vector of error terms on the parameters; all these variables are specified for the r'th iteration. $\nabla\underline{F}$ is the 1 x K gradient row vector of the first derivatives of the least-squares function, F $(\underline{\mu})$, defined at its optimal point where the gradient is zero.

Two possible schemes for solving equation (9.6) suggest themselves. If it is now assumed that the approximation to $\nabla\underline{F}$ is good enough using only first order information, then it is possible to assume that $\underline{G}_r = \underline{0}$, and make a further approximation of the following form

$$\underline{J}_r'\underline{f}_r' + \underline{J}_r'\underline{J}_r\underline{\varepsilon}_r' = \underline{0}\ . \tag{9.7}$$

Solving equation (9.7) for the error vector $\underline{\varepsilon}_r$ gives

$$\underline{\varepsilon}_r' = -(\underline{J}_r'\underline{J}_r)^{-1}\underline{J}_r\underline{f}_r'\ , \tag{9.8}$$

and equation (9.8) is then used as the basis of an iterative scheme in which new parameters $\mu_k(r+1)$ are chosen from $\underline{\mu}_{r+1} = \underline{\mu}_r + \underline{\varepsilon}_r$ until $\underline{\varepsilon}_r$ becomes less than some limit vector. The algorithm implied by equation

(9.8) is called Gauss's algorithm and it is clear that the system can be seen as equivalent to the notion of nonlinear regression. Indeed equation (9.8) has the same structure as the equation for computing the parameter values in a linear regression.

Of interest is the fact that if M = K, the system reduces to one in which the number of statistics is equal to the number of unknown parameter values (Kowalik and Osborne, 1968), then

$$\underline{\varepsilon}'_r = -\underline{J}_r^{-1} \underline{f}' \quad , \tag{9.9}$$

and equation (9.9) is clearly the Newton-Raphson equation which has been used quite extensively in conventional applications of these types of urban model (Batty, 1976). Gauss's algorithm would be linearly convergent in the neighbourhood of the optimum due to the fact that only first order information is used whereas if equation (9.6) were solved directly, the implied algorithm would be quadratically convergent. Then

$$\underline{\varepsilon}'_r = - (\underline{J}'_r\underline{J}_r + \underline{G}_r)^{-1}\underline{J}'_r\underline{f}'_r \quad , \tag{9.10}$$

and it is obvious that the iterative scheme based on equation (9.10) involves more computation than that in equation (9.8). One might also expect equation (9.10) to give better convergence than (9.8) but this involves the trade-off between speed and computer time per iteration which will be discussed quite fully in the sequel.

A much clearer interpretation of equation (9.10) can be made if a slightly different form of derivation is used. Consider the direct expansion of the least squares function $F(\underline{\mu})$ about some known value $F(\underline{\mu})_r$ on iteration r. Then using Taylor's theorem and expanding to terms of the second order leads to

$$F(\underline{\mu}) = \sum_k \frac{\partial F(\underline{\mu}_r)}{\partial \mu_k(r)} \varepsilon_k(r) + \tfrac{1}{2} \sum_{k\ell}\sum \varepsilon_k(r)\varepsilon_\ell(r) \frac{\partial^2 F(\underline{\mu}_r)}{\partial \mu_k(r)\partial \mu_\ell(r)} \quad , \qquad (9.11)$$

where $\varepsilon_k(r)$ is the error term associated with the parameter $\mu_k(r)$. As before assuming that the approximation in equation (9.11) is good, then equation (9.11) can be written in matrix terms as

$$F(\underline{\mu}) = F(\underline{\mu}_r) + \underline{\varepsilon}_r \nabla \underline{F}_r' + \tfrac{1}{2} \underline{\varepsilon}_r \underline{H}_r \underline{\varepsilon}_r' \quad , \qquad (9.12)$$

where $\underline{H}_r$ is the K x K matrix of second-order partial derivatives known as the Hessian (Simmonds, 1976). Equation (9.12) directly includes the second order information which when combined with the first order information given by equation (9.4) provides the necessary and sufficient conditions for a minimum. This is guaranteed if the Hessian matrix is positive definite, that is, if the third term on the right-hand-side of equation (9.12) is positive. At the minimum, the derivative of equation (9.12) with respect to the error vector $\underline{\varepsilon}_r$ must be equal to zero, that is

$$\underline{0} = \nabla \underline{F}_r + \underline{H}_r \underline{\varepsilon}_r' \quad , \qquad (9.13)$$

and the solution for $\underline{\varepsilon}_r'$ follows directly from equation (9.13)

$$\underline{\varepsilon}_r' = - \underline{H}_r^{-1} \nabla \underline{F}_r' \quad . \qquad (9.14)$$

In this instance, equation (9.14) forms the basis of the iterative scheme and the algorithm is referred to as Newton's method. This is quite different from Gauss's algorithm or the Newton-Raphson method although if the function is appropriately defined, the two methods can be the same.

To note the equivalence of Newton's method and the scheme given by equation (9.10), all that need be done is to find forms for $\nabla \underline{F}_r'$ and $\underline{H}_r$

in terms of the original deviation statistics $\underline{f}_m$. Then differentiating $F(\underline{\mu})$ to get $\nabla \underline{F}'_r$ leads to

$$\nabla F_k(r) = \sum_m f_m(r) \frac{\partial f_m(r)}{\partial \mu_k(r)} ,$$

and a second differentiation gives

$$H_{k\ell}(r) = \sum_m \left\{ f_m(r) \frac{\partial^2 f_m(r)}{\partial \mu_k(r) \partial \mu_\ell(r)} + \frac{\partial f_m(r)}{\partial \mu_k(r)} \frac{\partial f_m(r)}{\partial \mu_\ell(r)} \right\} .$$

Comparison with equation (9.4) demonstrates that

$$\nabla \underline{F}'_r = \underline{J}'_r \underline{f}'_r , \qquad \text{and} \tag{9.15}$$

$$\underline{H}_r = -(\underline{J}'_r \underline{J}_r + \underline{G}_r) . \tag{9.16}$$

Substituting for $\nabla \underline{F}'_r$ and $\underline{H}_r$ from equations (9.15) and (9.16) into (9.14) gives equation (9.10) which shows that Newton's method is equivalent, in this instance, to the second order Gaussian scheme. In the work reported below, the first order method based on equations (9.8) or (9.9) is referred to as Gauss's algorithm in contrast to the second order method of equation (9.10) or (9.14) which is Newton's algorithm. It is proposed to test both but from *a priori* considerations, it would appear that Newton's algorithm would be faster overall due to its quadratic convergence in the neighbourhood of the optimum. Moreover, the information provided by the Hessian is most useful in that global convergence can be assured through the criterion of positive-definiteness. Divergence of Newton's method away from the optimum in the early stages of the algorithm can also be avoided for if the Hessian indicates divergence, such divergence can be minimised by forcing the Hessian to be positive-definite. Such techniques are given by Himmelblau (1972) but they were not required in this case.

The application of Gauss's and Newton's algorithms to the models in this chapter involve matching the iterative structure of this model with the iterative structure of the algorithms. The function $F(\underline{\mu})$ was particularly straightforward in that the two parameters $\mu_1(r)$ and $\mu_2(r)$ are associated with the two mean statistics based on the predicted work trips and service demands respectively. These mean trip lengths on iteration r are defined as $\bar{C}(r)$ for work trips and $\bar{S}(r)$ for service demands, and their observed values at the base date are $\bar{C}$ and $\bar{S}$ as specified previously. At each iteration r, these means are

$$\bar{C}(r) = \sum_i e_i(r) \sum_j t_{ij}(r) d_{ij} / \sum_i e_i(r) \quad ,$$

$$= \sum_{ij}\sum T_{ij}(r) d_{ij} / \sum_{ij}\sum T_{ij}(r) \quad , \qquad \text{and} \qquad (9.17)$$

$$\bar{S}(r) = \sum_i e_i(r) \sum_j t_{ij}(r) \sum_k s_{jk}(r) d_{jk} / \sum_i e_i(r) \quad ,$$

$$= \sum_{jk}\sum S_{jk}(r) d_{jk} / \sum_{jk}\sum S_{jk}(r) \quad , \qquad (9.18)$$

where $S_{jk}(r)$ are the service demands from j to k and $T_{ij}(r)$ are the work trips from i to j as introduced above.

It is clear that the deviation statistics which form the basis of the least-squares function can now be stated for this problem as

$$f_1(r) = \bar{C}(r) - \bar{C} \ , \quad f_2(r) = \bar{S}(r) - \bar{S} \quad , \qquad (9.19)$$

and the elements of the Jacobian - the first derivatives of the functions in equation (9.19) are

$$J_{1k}(r) = \frac{\partial \bar{C}(r)}{\partial \mu_k(r)} \ , \quad J_{2k}(r) = \frac{\partial \bar{S}(r)}{\partial \mu_k(r)} \quad , \quad k = 1,2. \qquad (9.20)$$

The elements of the gradient vector $\nabla \underline{F}_k(r)$ and the Hessian matrix $\underline{H}_k(r)$ immediately follow from equations (9.19) and (9.20) and these are stated as

$$\nabla F_k(r) = [\bar{C}(r) - \bar{C}] \frac{\partial \bar{C}(r)}{\partial \mu_k(r)} + [\bar{S}(r) - S] \frac{\partial \bar{S}(r)}{\partial \mu_k(r)}, \quad k = 1,2, \quad (9.21)$$

$$H_{k\ell}(r) = [\bar{C}(r) - \bar{C}] \frac{\partial^2 \bar{C}(r)}{\partial \mu_k(r) \partial \mu_\ell(r)} + \frac{\partial \bar{C}(r)}{\partial \mu_k(r)} \frac{\partial \bar{C}(r)}{\partial \mu_\ell(r)}$$

$$+ [\bar{S}(r) - \bar{S}] \frac{\partial^2 \bar{S}(r)}{\partial \mu_k(r) \partial \mu_\ell(r)} + \frac{\partial \bar{S}(r)}{\partial \mu_k(r)} \frac{\partial \bar{S}(r)}{\partial \mu_\ell(r)}, \quad k=1,2. \quad (9.22)$$

The derivatives of the mean trip length functions are quite straightforward and have been stated generally by Evans (1971). They are not presented here, largely because their presentation would be somewhat lengthy and would contribute little to the central argument of this chapter. In the computer program developed for these algorithms, these derivatives are computed from their analytic forms although whether or not this can be done depends upon the model structure within which they are embedded. This is a problem which will be examined in the following section for all the elements of the two calibration algorithms have now been described, and it is now necessary to focus upon specific applications to the test problem.

ITERATIVE OPTIMISATION OF THE PSEUDO-DYNAMIC PROCESS.

In applying both Gauss's and Newton's algorithms, the quest is to find a set of parameters $\mu_1$ and $\mu_2$, for which $F(\underline{\mu}) = 0$ thus implying that $\bar{C}(r) = \bar{C}$ and $\bar{S}(r) = \bar{S}$ where r is being used now to denote the iteration on which equilibrium is reached. In previous models, the equilibrium form of the general model is tested inside the iterative schemes implied by the two algorithms, but in this context, the idea is to match the

iterative schemes implied by the algorithms with the matrix iterative analysis of the model itself. This is the same logic as matching the biproportional procedure to the model's iterations and in the integrated algorithm to be discussed below, both constraints and calibration will be matched in this way. As in the constraints procedures, the idea is to calibrate the model during the generative and distributive processes which build up the model's predictions, thus correcting the direction towards which the calibration is proceeding by feeding back information about the state of the system to the calibration process. A similar idea was used in Chapter 7 to calibrate a related model structure although in that instance, the incremental rather than cumulative form of the process was used and backtracking was necessary.

The question once again is whether or not the calibration procedure will converge. Convergence can only be assured if it can be demonstrated that the Hessian is positive-definite which would imply that the response surface generated by the model in terms of the least-squares function is strictly convex. Such a proof does not appear possible in general terms although in all the tests so far, the Hessian has been positive-definite. However, Evans' (1971) work with an unidimensional function implies that the surface is certainly unimodal, and probably convex in the region where one is most likely to search. In other words, convexity is not assured but in the area where the function varies most, it appears convex.

There are several different ways in which these algorithms can be matched to the model's iterative process and three types which are henceforth called 'Structures' have been applied. In Structure I, each model iteration

which consists of generating and distributing population from the input employment and service employment from population is considered to be quite separate. At the end of each iteration, the parameter error vector $\underline{\varepsilon}_r$ is computed using the information about trip lengths predicted solely during the iteration. In this structure, $\bar{C}(r)$ is computed from $t_{ij}(r)$ which is then used in the computation of $\bar{S}(r)$, but $\bar{C}(r)$ depends only upon $\mu_1(r)$, not $\mu_2(r)$ whereas $\bar{S}(r)$ depends on both. In other words, population depends upon employment generated from the previous iteration but not from the employment generated on the current iteration. The structure of activity relationships is sequential not simultaneous, although in the long term equilibrium, activities will be simultaneously related. In examining this structure, it is clear that the Jacobian matrix $\underline{J}_r$ is lower triangular, that is

$$\underline{J}_r = \begin{bmatrix} \frac{\partial \bar{C}(r)}{\partial \mu_1(r)} & 0 \\ \frac{\partial \bar{S}(r)}{\partial \mu_1(r)} & \frac{\partial \bar{S}(r)}{\partial \mu_2(r)} \end{bmatrix} ,$$

and thus its inversion is slightly easier than if it were full. Although this makes little difference for a 2 x 2 matrix, it does demonstrate that the Newton-Raphson method might be preferable to all others for large models with many sectors linked in this type of recursive or sequential way. For over-determined systems, of course, or for Newton's algorithm, this property is not important.

Perhaps the most important feature of this sequential ordering of computations relates to the fact that any simultaneity which might ultimately exist in the model, does not occur in any iteration. Thus, constraint

factors which are determined at the end of each iteration are assumed constant during an iteration and do not vary with respect to the parameters. As Evans (1971) has shown, if the constraint factors do vary with the parameters as is the case in the conventional static model in which the model and its constraints procedure is embedded into the calibration, large sets of simultaneous equations have to be solved to get the derivatives. In fact, in the Area 8 model, the derivatives were computed numerically to avoid this problem, but by formulating the model in the manner developed here, analytic derivatives can be used. These points also pertain to the other two methods used. In Structure II, the idea is to overlap the iterations and to use information from a previous iteration if this is the latest information available. In Structure I, the derivatives of $\bar{C}(r)$ with respect to the service centre parameter $\mu_2(r)$ is zero but with respect to the previous value of this parameter $\mu_2(r-1)$, it is positive. Thus the Jacobian matrix $\underline{J}_r$ now becomes

$$\underline{J}_r = \begin{bmatrix} \dfrac{\partial \bar{C}(r)}{\partial \mu_1(r)} & \dfrac{\partial \bar{C}(r)}{\partial \mu_2(r-1)} \\ \\ \dfrac{\partial \bar{S}(r)}{\partial \mu_1(r)} & \dfrac{\partial \bar{S}(r)}{\partial \mu_2(r)} \end{bmatrix} ,$$

and this implies the idea of a periodic update of information in terms of the relationship between employment from a previous iteration and interaction in a current iteration.

This notion of updating the information about the pattern of interaction on the system can be taken even further. In a sense, the organisation of the two activity generations and distributions in one iteration is fairly arbitrary and there is no reason why the relationship of population to employment to new population needs to reflect the computations associated

with an iteration. It is possible to overlap the iterations even more tightly than in Structure II, and to compute the error vector $\underline{\varepsilon}_r$ afresh after each activity has been generated and distributed. This idea is developed in Structure III: retaining the index r as characterising a conventional iteration, assume first that population $\underline{p}(r)$ has just been generated and distributed from employment $\underline{e}(r)$. Then a new error vector $\underline{\varepsilon}_r^1$ is computed using trip length derivatives of which the following Jacobian matrix $\underline{J}_r^1$ is an example

$$\underline{J}_r^1 = \begin{bmatrix} \dfrac{\partial\bar{C}(r)}{\partial\mu_1(r)} & \dfrac{\partial\bar{C}(r)}{\partial\mu_2(r-1)} \\ \\ \dfrac{\partial\bar{S}(r-1)}{\partial\mu_1(r-1)} & \dfrac{\partial\bar{S}(r-1)}{\partial\mu_2(r-1)} \end{bmatrix} .$$

From this information, two errors $\varepsilon_1^1(r)$ and $\varepsilon_2^1(r)$ are calculated: a new parameter value for $\mu_1(r)$ is not required as it is not needed until the next iteration. However $\mu_2(r)$ is calculated as $\mu_2(r) = \mu_2(r-1)+\varepsilon_2^1(r)$, and this is used in the calculation of service demands and employment.

After these services have been generated and distributed, a new error vector is again calculated from the new information of which the following Jacobian $\underline{J}_r^2$ is representative:

$$\underline{J}_r^2 = \begin{bmatrix} \dfrac{\partial\bar{C}(r)}{\partial\mu_1(r)} & \dfrac{\partial\bar{C}(r)}{\partial\mu_2(r-1)} \\ \\ \dfrac{\partial\bar{S}(r)}{\partial\mu_1(r)} & \dfrac{\partial\bar{S}(r)}{\partial\mu_2(r)} \end{bmatrix} .$$

At this point, a new parameter $\mu_1(r+1)$ is calculated as $\mu_1(r+1)=\mu_1(r)+\varepsilon_1^2(r)$. A new parameter for $\mu_2(r)$ is not required yet and it is clear that in

this process although both parameters could be changing after each activity has been dealt with, it is only relevant to change the parameter for which the next activity is being computed. Thus in the above scheme, $\varepsilon_1^1(r)$ and $\varepsilon_2^2(r)$ are pieces of information which are incidental to the process and are not actually used directly. In short, the information appearing in the Jacobians above (and of course the second-derivative functions which are not shown) is that which *directly* affects the distribution of activities.

In these three structures, the two algorithms are applied to response surfaces which are varying at each iteration. The degree to which each structure relates each iteration to the previous one, is however, a way of gradually moving from one response surface to the next. In the case of Structure I, each iteration is quite separate and each response surface is regarded as different from the previous one although it is generally assumed that the previous parameter values are good starting points for the new search. In Structures II and III, the links are much stronger, and in the case of Structure III, the closest possible link is used. In fact, the logic behind this updating of response surfaces in Structures II and III is not unlike the Gauss-Seidel method of matrix iterative analysis which is based on using information once created for the prediction of new information (Varga, 1962). Indeed, many of the techniques of this chapter depend upon the notion of using information once it has been predicted, in this way. Moreover, the fact that analytic derivatives are easier to compute for these types of sequential structure reduces computation time although it should be noted that this is only the case if the model is subject to locational constraints. If not, analytic derivatives for the simultaneous equilibrium are also possible. A diagrammatic illustration of these three

structures in relation to the model's iteration process is presented in Figure 9.5.

Three structures and two algorithms give rise to six possible applications of the unconstrained optimisation methodology. In general, it would appear that Newton's method would be more efficient than Gauss's, and that structures which were based on updating response surfaces from iteration to iteration would be preferable: that is, Structure III would be more efficient than Structure II and Structure I. Thus overall, Newton's method applied to Structure III would seem most efficient from purely theoretical considerations. However, Newton's method takes 47% longer per model iteration than Gauss's algorithm in terms of computer time, and this added time is almost entirely accounted for by the evaluation of the second derivatives. In contrast, there is hardly any difference in terms of computer time between the three structures: Structures I and II are almost identical and Structure III takes 3% longer than these to compute.

Thus the final choice of algorithm will rest on a trade-off between efficiency of the method and computer time taken. As in the constraints procedures given earlier, the efficiency of the method is measured by two statistics based on comparisons between successive iterations. These statistics are based on the change in the mean trip lengths from iteration to iteration, suitably normalised by their observed values. For the residential location model, the appropriate statistic $\eta^p(r)$ is defined as

$$\eta^p(r) = \frac{|\bar{C}(r) - \bar{C}(r-1)}{\bar{C}} ,$$

and for the service centre model, $\eta^s(r)$ is defined as

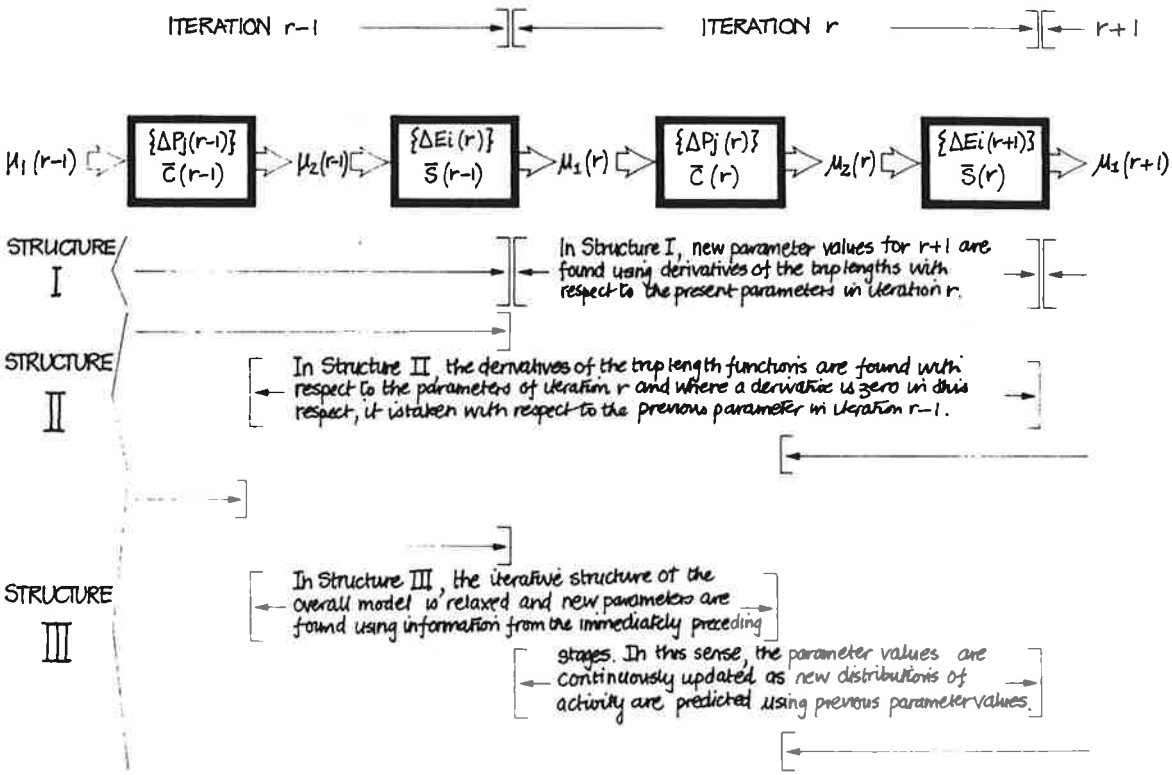



Figure 9.5: Calibration Structures.

$$\eta^s(r) = \frac{|\bar{S}(r) - \bar{S}(r-1)|}{\bar{S}} .$$

Appropriate convergence limits for these statistics have been fixed at $10^{-3}$ rather than $10^{-2}$ as in the case of the constraint factors. These statistics refer to the total system and a 0.1% change from iteration to iteration implies that for an observed mean trip length of 10 time/distance/cost units, precision to 2 decimal places has been obtained. In fact, the results presented below show that much greater accuracy can be obtained for calibration independent of the constraints procedure, and it may be that greater precision, up to $10^{-4}$ say, should be specified: this will clearly depend on the context and the resources available to build the model.

The results of the six applications - each algorithm applied to each structure - are recorded in Table 9.2 which shows the number of iterations required for each of the two statistics $\eta^p(r)$ and $\eta^s(r)$ to come within the specified limits. Clearly one of these statistics will dominate the convergence in that it will take longer for this statistic to converge, and the one that dominates in this way is picked out in bold type (underlined) in Table 9.2. The immediate impression from Table 9.2 is that there is little to distinguish the algorithms if the number of iterations to reach $10^{-7}$ is examined. However, there are some differences in terms of the speed of convergence before this limit is reached. The residential statistic converges faster at first up to the $10^{-3}$ limit but from then on the service statistic is faster. Up to $10^{-3}$, there is little question that Gauss's algorithm is faster for every structure but from then on until the limit of $10^{-7}$ is reached, there is little to choose, although Newton's algorithm probably has the edge. With regard to the three structures, there is hardly any differences in terms of speed although

Table 9.2: Convergence of the Calibration Procedures in Terms of the Numbers of Model Iterations Required to Reach Given Limits.

| Method* | Gauss's Algorithm | | | | | | Newton's Algorithm | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Convergence Limit | Structure I | | Structure II | | Structure III | | Structure I | | Structure II | | Structure III | |
| $10^{-1}$ | 3 | 2 | 3 | 2 | 3 | 2 | 5 | 2 | 5 | 2 | 5 | 2 |
| $10^{-2}$ | 5 | 3 | 5 | 5 | 5 | 5 | 8 | 3 | 8 | 4 | 8 | 4 |
| $10^{-3}$ | 6 | 5 | 6 | 6 | 6 | 6 | 9 | 7 | 9 | 8 | 9 | 8 |
| $10^{-4}$ | 10 | 11 | 10 | 11 | 10 | 11 | 9 | 11 | 10 | 10 | 10 | 10 |
| $10^{-5}$ | 14 | 15 | 14 | 15 | 14 | 15 | 14 | 15 | 14 | 15 | 14 | 15 |
| $10^{-6}$ | 19 | 19 | 19 | 20 | 19 | 20 | 19 | 19 | 19 | 20 | 19 | 20 |
| $10^{-7}$ | 23 | 25 | 23 | 25 | 23 | 25 | 24 | 25 | 23 | 25 | 25 | 25 |

* Note that in each column of the table, the number of iterations associated with the service statistic is presented first, then the number associated with the residential statistic.

in the case of Newton's algorithm, Structures II and III are marginally superior to Structure I. But given the relative crudeness of this type of comparison, this is hardly a basis for choice. In terms of the acceptable limit of $10^{-3}$, Gauss's algorithm is superior but in view of the general ambiguity of these results, it was decided to test all six methods in the integrated algorithm described below.

In results such as these, the basis for choice must be in terms of other factors and because Newton's algorithm takes 47% more computer time, Gauss's algorithm is to be preferred. In fact, the Newton-Raphson version of Gauss's algorithm has already been used extensively for models such as these, but the improvement in these applications over present practice is quite remarkable. In terms of Gauss's algorithm, only 6 model iterations are required to come to within the $10^{-3}$ limit and at this point, only just enough activity has been generated by the model in terms of its generative process. Using any of these methods, an excellent calibration can be achieved in 10 model iterations and this must be contrasted with something in the order of at least 70 in conventional applications (Batty, 1976). Moreover, this dramatic increase in speed results from a very different interpretation of the model's properties, and it also opens up the model structure to the incorporation of other information which might be considered important to the model's process. For example, although this has not been incorporated here, this model allows for the possibility that prior information affecting interaction can be used to start the model's distributive processes: the search for stable matrices of interaction need not be based on purely external criteria but on the idea that the model's pseudo-dynamic process is an approximation to the way the system has actually evolved and must be treated as such.

To conclude this section, it is worthwhile illustrating some detailed results from these tests so that a deeper impression of the speed of these various algorithms and their structure is given. In Figure 9.6 the two statistics of convergence are plotted on log-log graphs for Newton's algorithm applied to Structures II and III. The speed of convergence and directness of path in Structure II is clear and the stepped effect in Structure III is due to the fact that the convergence Statistics are computed each time the algorithm is applied which in this case is twice in each model iteration. In fact, Structures II and III have quite similar paths of convergence if the stepped effect is ignored. The detailed computations of intermediate variables in the algorithm such as derivatives are shown in Figure 9.7 for Newton's algorithm applied to Structure II. The signs of the first derivatives of the mean trip length functions are negative as expected from theory (see Evans, 1971), and the second derivatives of the least-squares function are also plotted. These partial derivatives are interesting in that the self-derivatives are some 100 times as large as the cross-derivative, thus implying that the Hessian is positive over the range of values shown.

Moreover, the first derivatives of the mean trip length functions with respect to each other's parameters are small relative to the same derivatives taken with respect to their own parameters, thus implying that the connections between the residential and service sectors of the model are much weaker than connections within these sectors. Finally, the well-known inverse relationship between parameter value and mean trip length is illustrated by Figure 9.7. The trip lengths converge to their observed values from above and the parameters from below and the graph also illustrates the relative speed of this convergence in terms of both the residential and service sectors. At this point, it

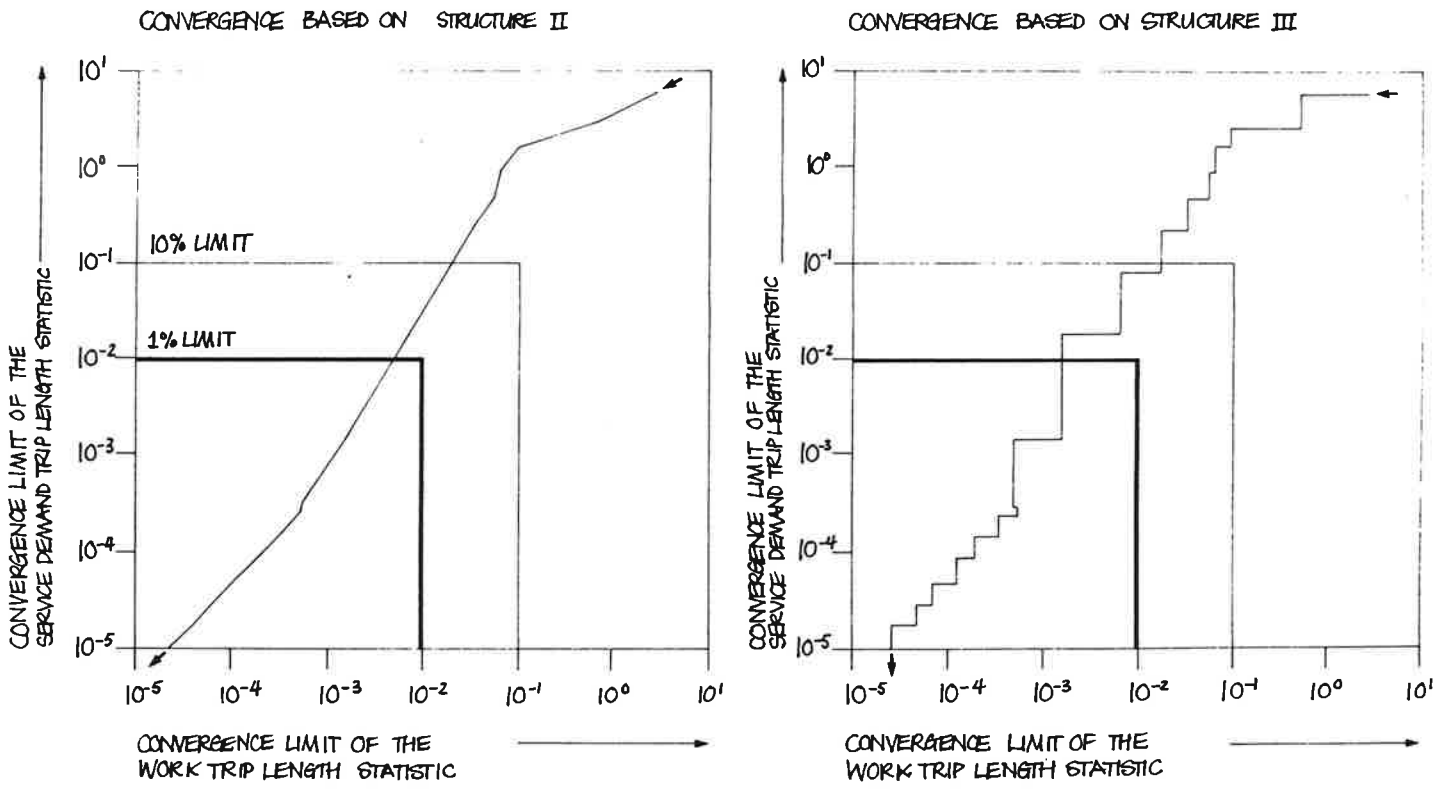



Figure 9.6: Convergence of the Newton Method on Structures II and III.

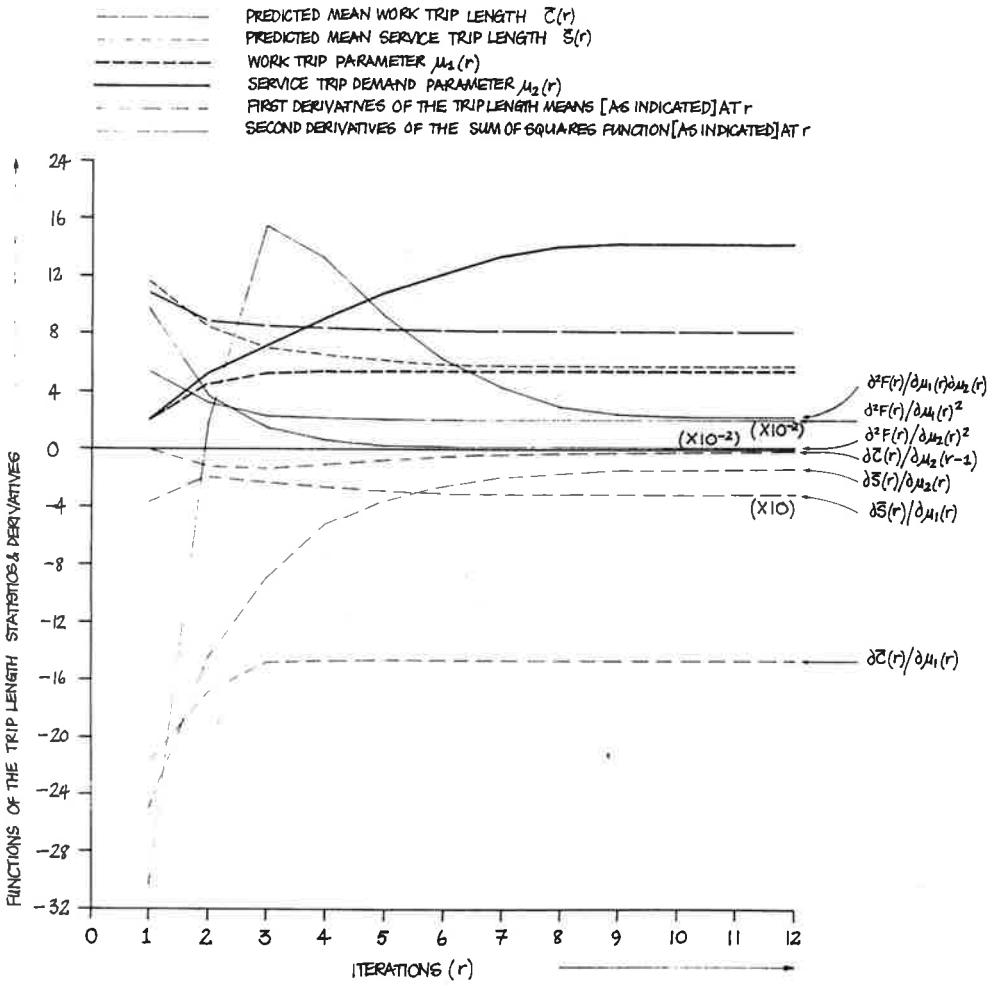



Figure 9.7: Convergence of the Trip Length Functions Using Newton's Method on Structure II.

is necessary to further test these calibration algorithms in the context of the various constraints procedures outlined above and to this end, the integrated algorithm will now be presented.

## AN INTEGRATED ALGORITHM FOR CONSTRAINED SOLUTION AND CALIBRATION.

From the test results already reported for the individual analyses of constraints and calibration procedures, it is not clear which methods are best, due to conflicting evidence in terms of the criteria used. Some indications are available that, for example, the polynomial methods appear best in terms of the constraints procedures, but when the two sets of methods - calibration and constraints - are both matched to the model's iterative structure, a new problem is constituted. It may be that the two procedures will reinforce one another or not, and thus it was decided to test a variety of combinations of methods already introduced on the general assumption that the results of the individual procedures could not be transferred to this problem involving joint procedures.

Three constraints procedures were developed: first, the original procedure, second, the procedure based on anticipating the model's activity generation over all iterations and third, the polynomial method based on anticipation over all iterations. The six methods of calibration based on the three structures and two algorithms outlined above were applied with each of these constraints procedures, making 18 applications in total. As in the calibration results, the essential difference in computer time for these applications related to the Gauss and Newton algorithms where the difference in computation time of 47% results from the need to evaluate second derivatives in Newton's algorithm.

The differences in terms of calibration structures and constraints procedures amounted to no more than 4% computer time per model iteration which is not significant.

The initial test runs of the integrated algorithm in which new parameters and factors were computed during or at the end of each model iteration for use in the next, were based on the test problem used in the individual analyses. However, it was eventually realised that the locational constraints specified for this problem were so tight that the observed mean trip lengths were not consistent with the range of possible solutions to the model. In other words, the constraints and trip lengths could not be simultaneously satisfied, and because of the way in which the two procedures operate, the biproportional procedure will always dominate the unconstrained optimisation algorithm. In this sense, the initial problem proved to be an excellent test problem in that it was used to continually refine and check the integrated algorithm in the hope that a solution might exist. However, the lack of a solution was realised when (in desperation!) the number of model iterations was increased to 100, and the results showed the parameter in the residential sector tending to infinity.

This is a particularly interesting result in models of this kind which in this case, is largely due to the relatively poor performance of the model on the test problem; it demonstrates that a set of constraints and trip lengths can be specified which in terms of the model's performance, it is not possible to meet. Furthermore, this problem could be particularly important in the context of using this type of structure in more formal problems of policy optimisation. It results from the

relationship between the various sectors which in the equilibrium are simultaneously related and it is certainly an area worthy of further research. An additional point also suggests itself: had both the constraints and calibration problem been set up as an unconstrained optimisation procedure in the manner suggested in the work of Robillard and Stewart (1974), the non-existence of a solution might never have been found for lack of convergence may then have been attributed to the method, and not the model as applied to this problem.

Because of these difficulties, a new test problem was defined by relaxing the locational constraints placed on the LTS data in such a way that a solution is ensured. The four statistics defined earlier in relation to the constraints procedure and the two in relation to calibration, were used to measure the convergence of each of the test runs. To summarise all this information, three tables will be presented for the limits $10^{-1}$, $10^{-2}$ and $10^{-3}$, and in each table, the maximum number of iterations required to meet the limit on the calibration statistics, $\eta^p(r)$ and $\eta^s(r)$, and the maximum on the constraint statistics, $\theta^p(r)$, $\theta^s(r)$, $\xi^p(r)$ and $\xi^s(r)$, will both be shown. The speed of convergence can be extracted by following through each of the three tables in sequence, and it is possible to note the differences in convergence of the calibration and constraints procedures. Table 9.3 shows the results for each of the 18 methods for the limit $10^{-1}$ and it is fairly clear from this, that Gauss's algorithm is faster than Newton's. In Table 9.4 which gives the number of iterations required to reach $10^{-2}$, Gauss's algorithm still has the edge. However, the picture begins to change in Table 9.5: for the limit $10^{-3}$, it is clear from Table 9.5 that Gauss's algorithm is still marginally better,

Table 9.3: Convergence of the Integrated Algorithm to the Limit $10^{-1}$

| Calibration Algorithms | Biproportional Procedure* | The Original 'Furness' Procedure | | The Procedure with Anticipated Activity Generated | | The Polynomial Procedure | |
|---|---|---|---|---|---|---|---|
| | Structure I | 3 | 21 | 3 | 18 | 3 | 21 |
| Gauss's Algorithm | Structure II | 3 | 21 | 4 | 18 | 4 | 18 |
| | Structure III | 3 | 21 | 4 | 18 | 4 | 18 |
| | Structure I | 5 | 21 | 5 | 18 | 9 | 15 |
| Newton's Algorithm | Structure II | 5 | 21 | 5 | 18 | 10 | 24 |
| | Structure III | 5 | 21 | 5 | 18 | 11 | 18 |

* Note that in Tables, 9.3, 9.4 and 9.5 the maximum number of iterations to reach the limit in terms of the two trip length statistics $\eta^P(r)$ and $\eta^S(r)$ is presented first in each column, followed by the maximum number in terms of the biproportional factor statistics $\theta^P(r)$, $\theta^S(r)$, $\xi^P(r)$ and $\xi^S(r)$.

Table 9.4: Convergence of the Integrated Algorithm to the Limit $10^{-2}$

| Calibration Algorithms | Biproportional Procedure | The Original 'Furness' Procedure | | The Procedure with Anticipated Activity Generated | | The Polynomial Procedure | |
|---|---|---|---|---|---|---|---|
| | Structure I | 5 | - | 5 | - | 10 | 28 |
| Gauss's Algorithm | Structure II | 4 | - | 5 | - | 11 | 28 |
| | Structure III | 5 | - | 5 | - | 10 | 28 |
| | Structure I | 8 | - | 8 | - | 13 | 27 |
| Newton's Algorithm | Structure II | 8 | - | 8 | - | 10 | 27 |
| | Structure III | 8 | - | 8 | - | 10 | 26 |

but after that point, that is after iteration 15 of the model, Newton's algorithm is clearly faster and is essential when Table 9.3 and 9.4 are analysed in terms of the constraints procedures.

In terms of the three structures, it is clear that Structure I is best up to the $10^{-1}$ limit but after that Structures II and III appear better. As previously, the evidence does not favour strong conclusions but Structure III probably has the edge. For the three constraints procedures, the evidence is however much firmer: up to the $10^{-1}$ limit, the anticipation method is best but after that the polynomial method is the only one which is able to meet the prespecified limit of $10^{-2}$. On balance for a large problem, Newton's method is probably superior and almost certainly the polynomial version of the biproportional constraints procedure will be required. The operation of this scheme using Structure II or III appears to give better results than on Structure I, and this conclusion is quite consistent with theoretical arguments which also suggest the same. By iteration 30, the Newton method applied to Structure III with the polynomial procedure is noticeably faster than any of the others and therefore, the results from this application will be presented in more detail.

Figure 9.8 shows the convergence of the biproportional factors themselves for the service and residential sectors and also contains a plot of the statistics $\theta^s(r)$, $\theta^p(r)$, $\eta^s(r)$ and $\eta^p(r)$. These graphs have many of the characteristics of the individual analyses presented earlier and although the order of magnitude of the biproportional factors has been established by the 15'th iteration, the polynomial method tends to perturb these values at every 4th iteration, thus

Table 9.5: Convergence of the Integrated Algorithm to the Limit $10^{-3}$.

| Calibration Algorithms | Biproportional Procedure | The Original 'Furness' Procedure | | The Procedure with Anticipated Activity Generated | | The Polynomial Procedure | |
|---|---|---|---|---|---|---|---|
| Gauss's Algorithm | Structure I | 11 | - | 8 | - | 26 | - |
| | Structure II | 11 | - | 8 | - | 11 | - |
| | Structure III | 11 | - | 8 | - | 11 | - |
| Newton's Algorithm | Structure I | 12 | - | 13 | - | 14 | - |
| | Structure II | 12 | - | 13 | - | 14 | - |
| | Structure III | 12 | - | 13 | - | 15 | - |

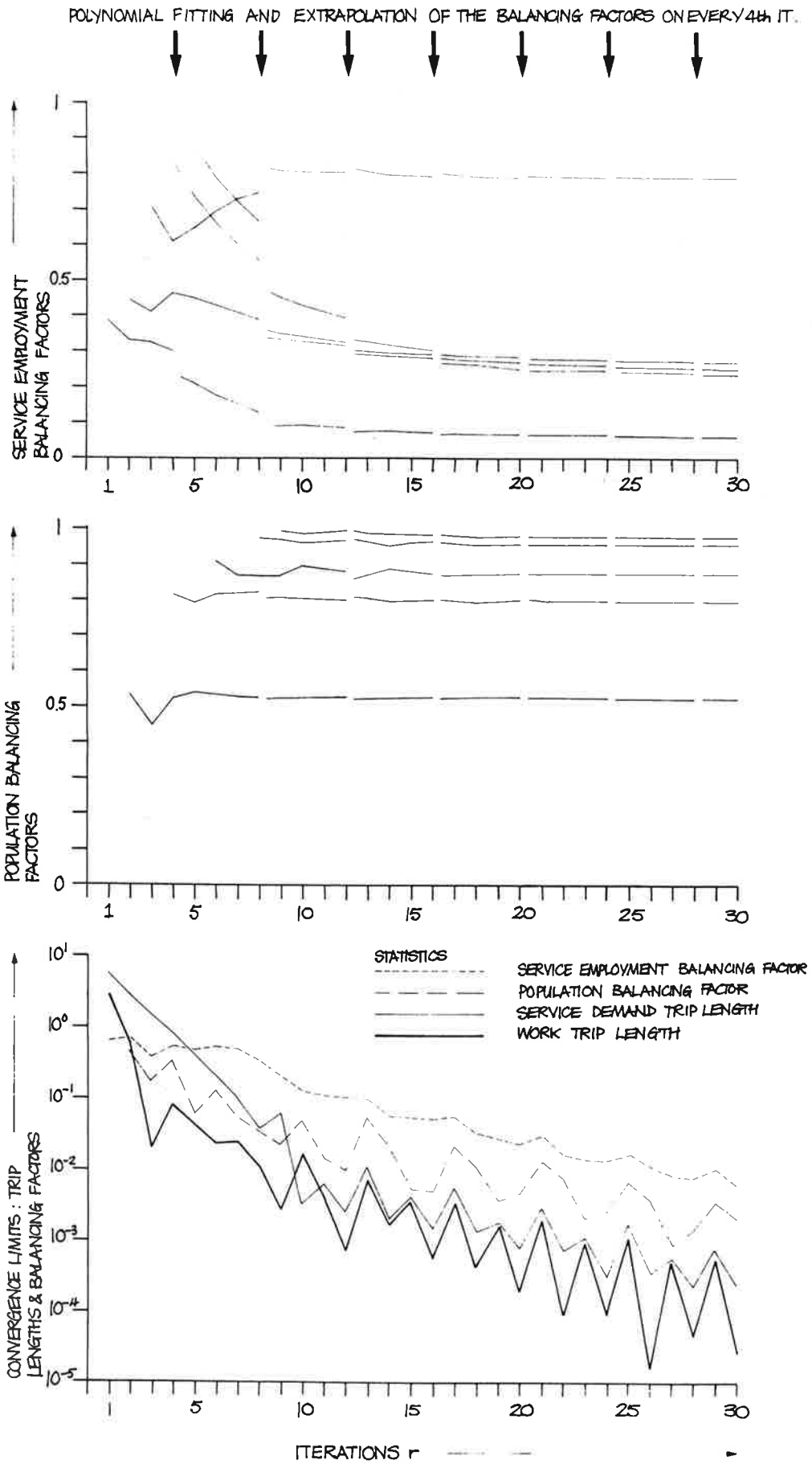



Figure 9.8: Convergence of the Integrated Algorithm Based on Newton-Structure III - Polynomial Methods.

accounting for the oscillations in the convergence statistics. Another important feature of the algorithm is the fact that up to about the 10'th iteration, the calibration algorithm converges at a similar speed to the same algorithm applied independently of the locational constraints procedure. But after this point, the convergence begins to slow and from about the 15'th iteration on, the calibration is dominated by the constraints procedure: both procedures converge at much the same speed. This is a particularly essential point in that the advantages of the calibration procedure tend to be lost in the later iterations, as the constraints procedure begins to dominate, and thus the algorithm as structured is highly dependent upon the constraints procedure. The implication must be that better and faster constraints procedures are required and this is clearly another important area for further research.

To formally establish the eventual speed of convergence, this particular version of the integrated algorithm has been run for 100 iterations and in Table 9.6 the long term convergence properties of the model are demonstrated by recording the limit reached on every 10'th iteration for each of the 6 statistics. This table shows that the convergence is fairly regular: in terms of the calibration statistics one additional decimal point of accuracy is gained after about every 15'th iteration after iteration 30, and for the biproportional factor and distribution statistics, this gain in precision is achieved after about every 20'th iteration. Overall in this example, it is possible to say that $10^{-1}$ accuracy is achieved within 20 iterations, $10^{-2}$ within 30, $10^{-3}$ within 50, $10^{-4}$ within 70 and $10^{-5}$ within 90. Clearly the convergence would be different for a larger problem, and by way

Table 9.6: Convergence of the Integrated Algorithm at Every 10'th Iteration Based on the Newton-Structure III Version.

| Convergence Statistics | Service Centre Location Submodel | | | Residential Location Submodel | | |
|---|---|---|---|---|---|---|
| Number of Iterations | $\eta^s(r)$ | $\theta^s(r)$ | $\xi^s(r)$ | $\eta^p(r)$ | $\theta^p(r)$ | $\xi^p(r)$ |
| 10 | $10^{-2}$ | $10^{0}$ | $10^{0}$ | $10^{-1}$ | $10^{-1}$ | $10^{-1}$ |
| 20 | $10^{-3}$ | $10^{-1}$ | $10^{-1}$ | $10^{-3}$ | $10^{-2}$ | $10^{-2}$ |
| 30 | $10^{-3}$ | $10^{-2}$ | $10^{-2}$ | $10^{-4}$ | $10^{-2}$ | $10^{-3}$ |
| 40 | $10^{-4}$ | $10^{-2}$ | $10^{-3}$ | $10^{-5}$ | $10^{-3}$ | $10^{-3}$ |
| 50 | $10^{-4}$ | $10^{-3}$ | $10^{-3}$ | $10^{-5}$ | $10^{-3}$ | $10^{-3}$ |
| 60 | $10^{-5}$ | $10^{-3}$ | $10^{-4}$ | $10^{-6}$ | $10^{-4}$ | $10^{-4}$ |
| 70 | $10^{-5}$ | $10^{-4}$ | $10^{-4}$ | - | $10^{-4}$ | $10^{-4}$ |
| 80 | $10^{-6}$ | $10^{-4}$ | $10^{-5}$ | - | $10^{-5}$ | $10^{-5}$ |
| 90 | - | $10^{-5}$ | $10^{-5}$ | - | $10^{-5}$ | $10^{-5}$ |
| 100 | - | $10^{-5}$ | $10^{-6}$ | - | $10^{-6}$ | $10^{-6}$ |

Note: The lack of a record in certain columns of the above table which is indicated by - implies that the relevant statistic had converged to $10^{-6}$ but no further convergence limit was specified in the computer print out which only gave results to 6 decimal places.

of conclusion to this analysis, the results of applying the integrated algorithm to the Central and West Berkshire model will now be presented.

For the 63 zone model of Berkshire, test runs were made for both the calibration procedure without locational constraints (unconstrained) and with such constraints. Furthermore, tests were made starting the matrix iterative solution of the model from the normal input, basic employment $\underline{e}(o) = \underline{b}$, and from the observed distribution of employment $\underline{e}(o) = \underline{\tilde{e}}$. A total of four runs were thus generated and the results are shown in Figure 9.9 for the statistics $[\eta^s(r)+\eta^p(r)]/2$ and $[\xi^s(r)+\xi^p(r)]/2$. It is clear that the convergence is generally slower than in the LTS test problem but by iteration 30, all these statistics had converged to the limit $10^{-1}$, and in the case of the calibration statistics to $10^{-3}$. In fact, in the unconstrained runs, the version starting with $\underline{e}(o) = \underline{\tilde{e}}$ converges to the limit $10^{-5}$ within 15 iterations for the calibration and within 18 iterations for the distribution statistics. This is in contrast to similar limits obtained within 20 iterations for calibration and 28 iterations for the distribution starting with $\underline{e}(o) = \underline{b}$.

Thus it is clear that to calibrate the model unconstrained, it is preferable to start with the observed vector of employment and to operate on this. Figure 9.9 also shows that this advantage is lost when the constrained version of the model is applied, for the constraints procedure tends to dominate the convergence, and the perturbations due to the polynomial smoothing become characteristic. The biproportional factor statistics, $\theta^s(r)$ and $\theta^p(r)$, are not shown

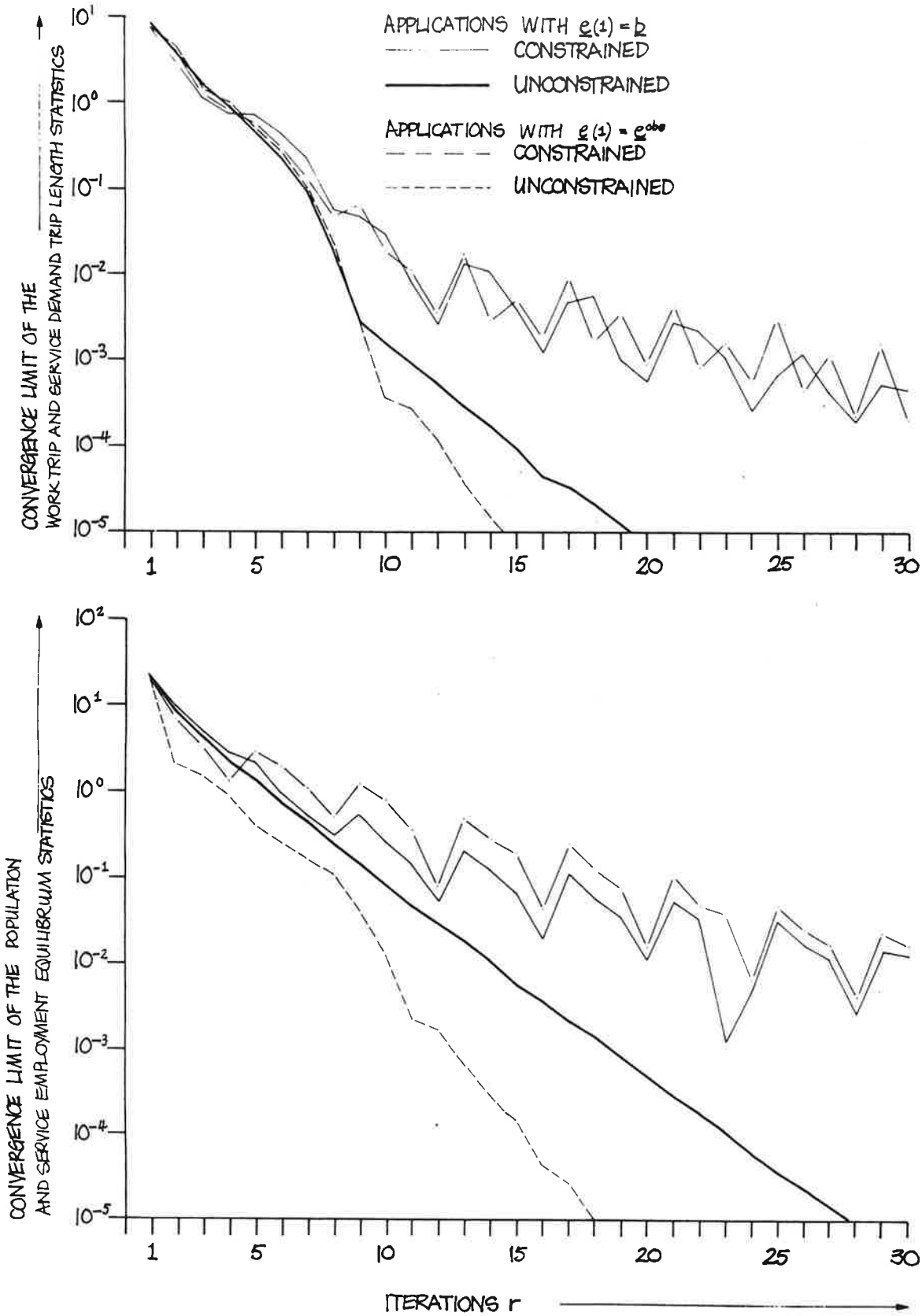



Figure 9.9: Convergence of the Central and West Berkshire Model.

in Figure 9.9 but these statistics also come within $10^{-1}$ in 30 iterations for the constrained runs and although this is not an acceptable limit, some runs to 50 iterations show that the limit of $10^{-2}$ is reached at about 38 iterations. This confirms that in large models, the integrated algorithm is far superior to any available technique for the fastest method previously reported required over 80 runs of a 34 zone model to reach a limit of $10^{-1}$ (Batty, Bourke, Cormode and Anderson-Nicholls, 1974). Yet this example once again reinforces the obvious point that better methods of dealing with locational constraints in such models are required if computer time is to be further reduced, and it is certain that further improvements can be made.

## CONCLUSIONS.

The main theme pursued in the last four chapters of this thesis has been the improvement of procedures for calibrating and constraining urban models through a reinterpretation of their equilibrium structure in dynamic terms. From the results presented in Chapters 7 and 8, it is quite obvious that enormous improvements in such procedures are possible, especially in the calibration process and to a lesser extent in methods for incorporating locational constraints. Although the methods which have been developed retain the equilibrium properties of the conventional static model, there is no reason why this should be so if it is hypothesised that the model's dynamic process is an approximation to the evolution of the system of interest, and the termination of the process treated accordingly. In particular, by reinterpreting such static models in dynamic terms, it is possible

to incorporate notions about prior and posterior events which influence the system, and this enables the overall model structure to include submodels based on the processes of dynamic information-minimising explored in earlier chapters.

The constraints procedure developed here based on biproportional factoring remains the most problematic element in the integrated algorithm and future research will surely have to be addressed to the search for faster methods. Although the use of methods of expansion or extrapolation based on assessing the interdependence of the system appear slow from the work of Robillard and Stewart (1974) and Cesario (1973), it may be possible to develop certain gradient procedures which anticipate the equilibrium values of the factors more closely than the polynomial method. Such techniques would not involve any matrix inversion which is the stumbling block with the method of Robillard and Stewart, although other problems such as the way in which these methods would fit into the matrix iterative analysis would have to be faced. It is also possible that entirely different sorts of procedures might be more appropriate to this problem. Scarf's (1973) algorithm, for example, which is used to generate solutions to the general economic equilibrium model, is based on systematic sampling of the possible solutions according to certain known properties of the solution space, and it does appear that such methods might be relevant to the problem.

Two other avenues for major research immediately arise from the ideas presented in this chapter. The first relates to the development of matrix iterative analysis and its correspondence with the complete

mover model. In using such a method, all activity is regenerated and redistributed through the model's iterations, and the relatively tractable form of this type of model suggests that it could provide a useful basis for fully-dynamic modelling. In moving from one state of the system to the next, a major component of the prediction is the existing distribution, and the complete mover model thus contains a large autoregressive term. Potentially this type of framework with its implied equilibrium properties could be adapted to the simulation of a real historical process, and the treatment of inert activity (stayers) could be fashioned by partitioning the distribution matrices into mover-stayer components. In a sense, this type of approach itself is based on updating a system using all available information, and conceptually, it is consistent with the ideas of information-minimising. The second possibility for future research involves the general question of calibration and constraint. It was hinted in the discussion that both problems could be treated in a more integrated fashion as a problem in optimisation and there could be further improvement to conventional models in terms of speed. Moreover, a more formal treatment of the optimal policy question could be developed in this context, and the idea of matching the model's structure with the iterative nature of such optimisation is appealing.

Finally, the last four chapters have introduced an idea which hopefully might be of more general importance in building operational urban models. Simultaneous systems with well-defined equilibrium properties such as that on which the conventional static model is based, can be opened up and treated sequentially without necessarily destroying their equilibrium properties. Disaggregation of such

models and their extension to incorporate other sectors should be considered in terms of their implications for dynamic processes, for if the same kind of sequential structures can be developed, such models are clearly more feasible in operational terms. Indeed in larger models, such as those indicated in Chapter 2, such reformulations and reinterpretations might make the difference between their practical application or not. And for this reason as well as those pertaining to a better theoretical understanding, it appears that continued work on this approach is necessary.

# CHAPTER 10.

## MARKOV PROCESSES IN LINEAR URBAN MODELS.

The arguments developed so far in this thesis are based on the general notion that the multiplier effects contained within urban models of the Lowry type can be utilised for a variety of purposes: clearly to enable such models to be elaborated in a temporal context (Batty, 1976) but of greater importance here, to enable such models to be solved in the most efficient way. To achieve these purposes, the models and methods developed in earlier chapters involved some rather extensive and somewhat laborious algebraic elaborations, and so far, there has been little which links these models to the conventional analysis of mathematical dynamics. However in the presentation of the algorithm for adaptive calibration in Chapters 7 and 8, it was clear that the traditional urban model's multiplier processes could be examined in more formal terms. In fact, the property of separability in which the economic base and spatial interaction components of such models are independent suggests that these multiplier processes can be explored as the conjunction of two processes: as a converging economic base mechanism and as a spatial averaging process.

The properties of the economic base mechanism are well known but

that of spatial averaging has not been investigated. However in Chapter 7, the spatial averaging process was examined in terms of changing trip lengths during the multiplier process. In Appendix 2, it was shown that this process was Markovian in form, and although the analysis was not taken very far, there was a tacit suggestion that a more formal analysis of spatial averaging would be fruitful. Such formal analysis has increased as the field has matured (Wilson, Rees and Leigh , 1977) and during this research, such analysis was given added impetus in 1978 by Schinnar. In his paper, Schinnar (1978) argued that models of the Garin-Lowry type (Garin, 1966; Lowry, 1964) can display a condition of locational invariance where non-basic employment is spatially independent of the basic employment which generates it. In short, Schinnar shows empirically that it is the model's distributing processes which lead to such invariance, and the consequences of his result imply that predictions from such models might be trivial.

This chapter will be concerned with generalising Schinnar's ideas and linking them to the formal analysis implied by Chapters 6 and 7 and briefly indicated in Appendix 2. Here the analysis will be short and sharp in contrast to earlier chapters, and this chapter in particular will develop the analysis from the standpoint of Schinnar's (1978) paper. In the next and final substantive chapter, the ideas of this chapter will be generalised and developed empirically, coming back full circle to many of the ideas raised in Chapter 2. In this chapter, the notation will differ a little from that in previous chapters in that the emphasis in the first instance at least, will be on relating these arguments to Schinnar's (1978) paper.

In hindsight, Schinnar's result is elementary for it suggests that if the interaction patterns linking population to employment and/or employment to population display a condition of distributional invariance, then activity patterns derived as a summation of interaction will also display this invariance. Schinnar demonstrated his result heuristically by constructing an invariant distributional pattern using some fairly realistic assumptions about the spatial variation of population and the demand for non-basic employment. It is however possible to go further and to interpret his result more formally; in fact to show that all models of the Garin-Lowry type and perhaps a wider class of models with an equivalent equilibrium structure, certainly the pseudo-dynamic models developed here, are subject to Schinnar's condition of invariant distributional regularity.

This chapter is concerned with presenting such a formal extension and with showing that distributional invariance is a natural consequence of the model structure. Moreover, it is possible to explore such invariance analytically and to derive measures for assessing the amount of invariance in any particular empirical application. Accordingly, the equilibrium structure of the Garin-Lowry model is first introduced and its form as a Leontief series is used to engender a formal analysis of its distributional and generative processes. In essence, the analysis reveals that sequential distribution in this model is characterised by the Markov property, and this clearly accounts for the invariance. This invariance is then investigated formally by concentrating on the eigenstructure of the distributional process. A series of measures which detect convergence to the invariant distribution are then postulated, and the theoretical results

produced are finally given substance by reworking the example used by Schinnar (1978), originally taken from Rogers (1971). Many insights emerge from the analysis, and these are drawn together in the conclusion which suggests directions to be explored in the penultimate chapter.

EQUILIBRIUM STRUCTURE OF THE GARIN-LOWRY MODEL.

In the model, population depends on basic and non-basic employment, and non-basic employment on population in two ways: economic base ratios link the absolute amount of population to employment and non-basic employment to population, and distribution rules map the set of workplaces into residences and residences into service centres. Basic employment provides the activity input to the model and population and non-basic employment are simultaneously related in the linear manner first noted by Harris (1966).

Using Schinnar's (1978) notation, population is derived from employment by

$$\underline{w} = \underline{A}^{-1}\underline{P}\,\underline{e} \,, \qquad (10.1)$$

where $\underline{w}$ is an n x 1 vector of population, $\underline{e}$ is an n x 1 vector of employment, $\underline{P}$ is an n x n journey from work-to-home transition probability matrix normalised so that $\underline{1}' = \underline{1}'\underline{P}$ ($\underline{1}$ is a n x 1 unit vector), and $\underline{A}$ is an n x n diagonal matrix of activity rates, that is, employment per capita. Non-basic employment is related to population in an analogous fashion

$$\underline{e}_{nb} = \underline{Q}\,\underline{B}\,\underline{w} \,,$$

$$= \underline{Q}\,\underline{B}\,\underline{A}^{-1}\,\underline{P}\,\underline{e} \,. \qquad (10.2)$$

where $\underline{e}_{nb}$ is an n x 1 vector of non-basic employment, $\underline{B}$ is an n x n diagonal matrix of population-serving ratios and $\underline{Q}$ is an n x n demand from home-to-service centre transition probability matrix normalised so that $\underline{1}' = \underline{1}'\underline{Q}$.

Two characteristics of the assumed equilibrium contained in equations (10.1) and (10.2) should be noted. First, the generation of activities is separate from their spatial distribution to the n zones or points of the spatial system. Thus the generative and distributive processes are independent and sequenced in the most obviour way. Second, the diagonal matrices $\underline{A}$ and $\underline{B}$ are scalar matrices in that $A_{ii} = \alpha, \forall_i$ and $B_{ii} = \beta, \forall_i$. This assumption of spatial uniformity in the activity and population-serving ratios ensures that the total activity - population and employment generated by the model, is constant no matter what distribution is implied by $\underline{P}$ and $\underline{Q}$. In most applications of this model, this assumption of constancy has been adopted (Batty, 1976) thus allowing for parametric adjustment to be carried out on $\underline{P}$ and $\underline{Q}$. In Schinnar's (1978) analysis, this assumption is not made but in this chapter, it is essential to the subsequent analysis, thus implying that the formal argument developed here is a little less general than that used by Schinnar. These assumptions enable equation (10.2) to be simplified to

$$\begin{aligned} \underline{e}_{nb} &= \underline{B}\,\underline{A}^{-1}\underline{Q}\,\underline{P}\,\underline{e}\ , \\ &= \mu\,\underline{Z}\,\underline{e}\ , \end{aligned} \tag{10.3}$$

where $\mu$ is the scalar formed from $B_{ii}A_{ii}^{-1} = \beta\alpha^{-1}$ and $\underline{Z}$ is a transition probability matrix incorporating the effects of the journey-to-work and service demand matrices. $\underline{Z} = \underline{Q}\,\underline{P}$ and it is thus clear that $\underline{1}' = \underline{1}'\underline{Z}$ for the product of two stochastic matrices is a stochastic matrix.

The equilibrium structure can now be easily derived. Using equation (10.3) in the identity

$$\underline{e} = \underline{e}_{nb} + \underline{e}_b \ , \tag{10.4}$$

where $\underline{e}_b$ is an n x 1 vector of basic employment, total employment is calculated from

$$\underline{e} = \mu \underline{Z} \underline{e} + \underline{e}_b \ , \tag{10.5}$$

If the equation system given by (10.5) has a solution, then equation (10.5) can be written

$$\underline{e} = (\underline{I} - \mu \underline{Z})^{-1} \underline{e}_b \ , \tag{10.6}$$

where $\underline{I}$ is an n x n identity matrix. There are several interpretations of equations (10.5) and (10.6). The simultaneous nature of the solution is self-evident (Harris, 1966) but it is also possible to interpret the solution in sequential terms. The inverse $(\underline{I} - \mu \underline{Z})^{-1}$ inequation (10.6) can be expanded and interpreted as the multiplier effects of an impact on the economic system (Garin, 1966) as in earlier chapters or the system of equations in (10.5) can be solved sequentially starting with an estimate for $\underline{e}$ on the RHS of equation (10.5) and iterating; this is the so-called Jacobi split solution method developed in Chapter 8. In both cases, equation (10.6) can be written in series form as

$$\underline{e} = \lim_{m \to \infty} (\underline{I} + \mu \underline{Z} + \mu^2 \underline{Z}^2 + \mu^3 \underline{Z}^3 + \ldots + \mu^m \underline{Z}^m) \underline{e}_b \ ,$$

$$= \sum_{m=0}^{\infty} \mu^m \underline{Z}^m \underline{e}_b , \tag{10.7}$$

where $\mu^0 = 1$ and $\underline{Z}^0 = \underline{I}$. The property of distributional invariance depends upon exploiting this series expansion of the solution. This

is only consistent with equation (10.6) if the series converges; this will be the case if the dominant eigenvalue of $\mu\underline{Z} < 1$ which can only occur if $\mu < 1$. These considerations will be explored later. At present, it is sufficient to note the fact that $\mu$ controls the absolute amount of activity generated and $\underline{Z}$ its distribution. Thus a study of distributional invariance should initially begin with a study of $\underline{Z}$.

SPATIAL DISTRIBUTION AS A MARKOV PROCESS.

Consider a model based on equation (10.7) in which there are no limits placed on the value of $\mu$. If $\mu < 0$, the model simulates an explosive oscillation which has no obvious meaning in terms of urban phenomena. But if $\mu < 0 < 1$, the model can be used to simulate the conventional multiplier process associated with an initial impact $\underline{e}_b$ and the series converges to equation (10.6). If $\mu > 1$, the model predicts exponential growth with no finite limit which may be useful in simulating growth in the shorter term. However, it is instructive to examine the case of constant linear growth where $\mu = 1$. In such a case, the generative process becomes trivial and the emphasis is solely on distribution. From equation (10.7), each increment of non-basic employment generated on iteration m is given by the n x 1 vector $\underline{e}_{nm}(m)$ which is predicted from

$$\underline{e}_{nm}(m) = \underline{Z}^m \underline{e}_b \ . \qquad (10.8)$$

In effect, at each iteration of the model an amount of non-basic employment equivalent to the original input of basic employment is being added to the system.

The sequential process implied by equation (10.8) is a first order finite Markov chain for $\underline{Z}$ is a stochastic matrix. Thus it is worth examining the form of the distribution matrix $\underline{Z}^m$ as m becomes large. For a spatial system which is connected, $\underline{Z}$ will be a positive or non-negative matrix, and if non-negative, some power of $\underline{Z}$ will be positive. Thus $\underline{Z}^m$ will converge to the steady state matrix $\tilde{\underline{Z}}$ which is idempotent

$$\tilde{\underline{Z}} = \lim_{m \to \infty} \underline{Z}^m . \tag{10.9}$$

It is a standard result of Markov chain theory that each column of $\tilde{\underline{Z}}$ is equal to the n x 1 vector $\underline{x}_1$ which can be calculated by solving the system of equations

$$\underline{Z}\,\underline{x}_1 = \underline{x}_1 . \tag{10.10}$$

This system of equations in (10.10) is not linearly independent for it is easy to show that the rank of $\underline{Z} \leq n - 1$. Assuming that the rank is n - 1, then $\underline{x}_1$ can be calculated if an additional linearly independent equation for $\underline{x}_1$ is available. As $\underline{Z}$ is stochastic and $\underline{x}_1$ is to be used as a column of $\tilde{\underline{Z}}$, it is usual to solve for $\underline{x}_1$ taking the first n - 1 equations from (10.10) together with the normalisation equation $\underline{1}'\underline{x}_1 = 1$ (Bailey, 1964).

Using these results, the limit of equation (10.8) can now be written in terms of the steady state matrix $\tilde{\underline{Z}}$. Then

$$\lim_{m \to \infty} \underline{e}_{nm}(m) = \tilde{\underline{Z}}\,\underline{e}_b ,$$

$$= \underline{x}_1\underline{1}'\underline{e}_b = E_b\underline{x}_1 , \tag{10.11}$$

where $E_b$ is the total basic employment in the system formed from $\underline{1}'\underline{e}_b$. Equation (10.11) demonstrates that the distribution of non-basic employment in the limit is independent of the distribution of basic

employment, and in the model in which $\mu = 1$, this independent distribution of non-basic employment will eventually dominate the system. Furthermore, the property of invariance is a characteristic of the steady state matrix $\underline{\tilde{Z}}$. In fact, this result is not entirely unexpected for as more and more non-basic employment is generated in the sequence, its dependence upon the initial distribution of basic employment must lessen. The fact that the distribution matrix $\underline{Z}^m$ becomes distributionally invariant is of greater consequence for it can be argued that more plausible and more efficient models can be designed in which $\underline{Z}$ is non-stationary, for example, is a function of m. These types of models were discussed earlier in Chapters 3 to 9.

Returning to the conventional model based on equation (10.7) in which $0 < \mu < 1$, the Markov property and convergence to an invariant distribution still hold but their importance is clearly affected by the value of $\mu$. Because this model strictly separates generation from distribution, it is possible to interpret the model's structure as one which matches a generative process which is geometrically convergent in the absolute sense with a distributive process which is geometrically convergent in the relative sense. The importance of the invariant distributional property will thus depend upon the interaction between the economic base mechanism and the spatial Markov process and thus there emerge a number of ways in which the model's predictions tend towards an invariant distribution. The economic base ratio $\mu$ is the ratio of non-basic to total employment, that is, $\mu = \beta\alpha^{-1} = \underline{1}'\underline{e}_{nb}/\underline{1}'\underline{e}$. Then as $\mu \to 0$, $\underline{e}_{nb} \to \underline{e}$ and $\underline{e}_b \to 0$; in such a case, there is more and more non-basic employment to generate

and thus in absolute terms, invariance can become more important. In terms of distribution, the closer $\underline{Z}$ is to its steady state matrix $\underline{\tilde{Z}}$, the more important the distributional invariance. It is therefore clear that the same degree of importance associated with the invariance property can result from different values of $\mu$ and $\underline{Z}$. Because the model is based on a well-defined matrix representation, it is possible to formally explore these questions further, and to devise measures which show how important the invariance property is.

ANALYSIS OF INVARIANT DISTRIBUTIONAL REGULARITIES.

Consider the case where the distribution matrix $\underline{Z}$ is already in the steady state, that is, $\underline{Z}^m = \underline{Z} = \underline{\tilde{Z}}$, $m > 0$. Equation (10.7) now becomes

$$\underline{e} = \lim_{m \to \infty} (\underline{I} + \mu \underline{\tilde{Z}} + \mu^2 \underline{\tilde{Z}} + \mu^3 \underline{\tilde{Z}} + \ldots + \mu^m \underline{\tilde{Z}}) \underline{e}_b ,$$

$$= \underline{e}_b + \lim_{m \to \infty} (\mu + \mu^2 + \mu^3 + \ldots + \mu^m) \underline{\tilde{Z}} \underline{e}_b . \qquad (10.12)$$

Assuming $0 < \mu < 1$ henceforth, then equation (10.12) can be simplified further to

$$\underline{e} = \underline{e}_b + \frac{\mu}{1 - \mu} \underline{x}_1 \underline{1}' \underline{e}_b ,$$

$$= \underline{e}_b + \frac{\mu}{1 - \mu} E_b \underline{x}_1 . \qquad (10.13)$$

The scalar $\mu(1 - \mu)^{-1} E_b$ is the total amount of non-basic employment to be allocated and equation (10.13) clearly demonstrates its independence from the distribution of basic employment. Furthermore as $\mu \to 1$, $\underline{e}_b \to 0$ and this also reinforces the conclusion that distributional invariance can become critical if basic employment is small relative to non-basic. This result is consistent with one of Schinnar's (1978)

conclusions. To gauge the importance of invariance in any particular application, it would be necessary to compare the invariant distribution given by equation (10.13) with its real distribution predicted from equation (10.7). To do this, a form for $\underline{Z}$ is required in terms of its steady state vectors $\underline{x}_k$ and before this can be done, some results from matrix theory must be stated. These results were also summarised in Chapter 2 and in Appendix 2.

As $\underline{Z}$ is a stochastic matrix, there exists a similarity transformation which diagonalises $\underline{Z}$ into its eigenvalues and eigenvectors, that is, into its roots and their bases. Then

$$\underline{Z}\,\underline{X} = \underline{X}\,\underline{\Lambda}\ , \tag{10.14}$$

where $\underline{X}$ is an n x n matrix composed of the RH eigenvectors $\underline{x}_k$, k = 1,2,...,n, and $\underline{\Lambda}$ is an n x n diagonal matrix of eigenvalues $\lambda_k$ associated with their appropriate eigenvectors. The diagonalisation can be achieved immediately from equation (10.14).

$$\underline{Z} = \underline{X}\,\underline{\Lambda}\,\underline{X}^{-1}. \tag{10.15}$$

In a similar fashion, $\underline{Z}$ can be structured in terms of its LH eigenvectors $\underline{y}'_k$ , k = 1, 2,..., n which are ordered in an n x n matrix $\underline{Y}'$. Then

$$\underline{Y}'\underline{Z} = \underline{\Lambda}\,\underline{Y}'\ , \tag{10.16}$$

and the diagonalisation is given as

$$\underline{Z} = (\underline{Y}')^{-1}\,\underline{\Lambda}\,\underline{Y}' \tag{10.17}$$

Without loss of generality, it is assumed that all the eigenvalues are distinct and thus the LH and RH eigenvectors are linearly independent and orthogonal. A suitable choice of scalars is able to ensure that

$$\underline{X}\,\underline{Y}' = \underline{I}\ , \tag{10.18}$$

and these eigenvectors thus form an orthonormal basis for the diagonal representation of the matrix $\underline{Z}$ (Heal, Hughes and Tarling, 1974). From equation (10.18) it is immediately clear that $\underline{Y}' = \underline{X}^{-1}$ and $\underline{X} = (\underline{Y}')^{-1}$ and these results imply that $\underline{Z} = \underline{X}\,\underline{\Lambda}\,\underline{Y}' = (\underline{Y}')^{-1}\,\underline{\Lambda}\,\underline{X}^{-1}$. Note also that the LH and RH eigenvectors are arbitrary in the sense that the rank of $\underline{Z} \leq n - 1$ and must thus be normalised to ensure solution. The form adopted in equation (10.10) where $\underline{1}'\underline{x}_k = 1$ is convenient but any other normalisation is possible.

Using these diagonalisations, the matrix $\underline{Z}$ can be factored into distinct components associated with its eigenvalues and eigenvectors. Then

$$\underline{Z} = \sum_{k=1}^{n} \lambda_k\, \underline{x}_k\, \underline{y}_k'\ , \tag{10.19}$$

and this so-called spectral decomposition of the matrix $\underline{Z}$ (Bailey, 1964) gives a convenient form for computing powers of $\underline{Z}$

$$\underline{Z}^m = \sum_{k=1}^{n} \lambda_k^m\, \underline{x}_k\, \underline{y}_k'\ . \tag{10.20}$$

To explore this question further, the properties of stochastic matrices must be noted. Applying the Perron-Frobenius theorem to $\underline{Z}$ shows that $\underline{Z}$ has a dominant eigenvalue equal to 1 and that all other eigenvalues of $\underline{Z}$ are less than 1 in absolute value. If all the eigenvalues can be assumed to be distinct, the diagonalisation outlined above and the spectral decomposition in equations (10.19) and (10.20) are well-defined. In fact, the assumption of distinct eigenvalues is not particularly strong in this context for $\underline{Z}$ can be slightly perturbed to get rid of multiple roots without radically altering the model's predictions. Thus if it is assumed that the eigenvalues and vectors

are ordered so that $\lambda_1 > |\lambda_2| > |\lambda_3| > \ldots > |\lambda_n|$, then the limit of $\underline{Z}^m$ depends entirely upon $\lambda_1$; that is

$$\lim_{m \to \infty} \underline{Z}^m = \tilde{\underline{Z}} = \lambda_1 \underline{x}_1 \underline{y}_1',$$

$$= \underline{x}_1 \underline{1}', \tag{10.21}$$

where $\lambda_1 = 1$, the vector $\underline{x}_1$ is normalised to sum to 1, and $\underline{y}'_1 = \underline{1}'$ due to the fact that $\tilde{\underline{Z}}$ is a stochastic matrix.

Convergence of the conventional model in equation (10.7) can now be formally demonstrated. This convergence will clearly depend upon the difference between $\underline{Z}^m$ and $\tilde{\underline{Z}}$ which from equation (10.21) depends on the eigenvalues and vectors other than those associated with the dominant one. Then

$$\underline{Z}^m - \tilde{\underline{Z}} = \sum_{k=2}^{n} \lambda_k^m \underline{x}_k \underline{y}'_k \ . \tag{10.22}$$

Using the spectral decomposition in equations (10.19) and (10.20), equation (10.7) can now be rewritten as a function of the steady state and its convergence to that state. The first line of (10.7) is repeated for convenience

$$\underline{e} = \lim_{m \to \infty} (\underline{I} + \mu \underline{Z} + \mu^2 \underline{Z}^2 + \mu^3 \underline{Z}^3 + \ldots + \mu^m \underline{Z}^m) \underline{e}_b \ .$$

$$= (\underline{I} + \sum_{m=1}^{\infty} \mu^m \sum_{k=1}^{n} \lambda_k^m \underline{x}_k \underline{y}'_k) \underline{e}_b \ ,$$

$$= \underline{e}_b + \frac{\mu}{1 - \mu} E_b \underline{x}_1 + \sum_{m=1}^{\infty} \mu^m \sum_{k=2}^{n} \lambda_k^m \underline{x}_k \underline{y}'_k \underline{e}_b . \tag{10.23}$$

Equation (10.23) is a new representation of the Garin-Lowry model in

terms of its steady state, and it is clear that the last term on the RHS of (10.23) represents the distributive differential between the actual final state and the hypothetical steady state. The smaller this differential, the closer the model's predicted distribution to the invariant distribution. Thus the importance of the Markov property - the invariant distribution, can be measured in any particular application using statistics which compare equation (10.13) with equation (10.23).

MEASUREMENT OF THE INVARIANCE PROPERTY.

Three measures will be suggested here, the first emphasising only the degree of distributional invariance, the second and third relating this invariance to the absolute amount of activity being predicted. Consider first the actual increment of non-basic employment associated with the sequential process in equation (10.7). Then this increment can be written in terms of its spectral decomposition as

$$\underline{e}_{nb}(m) = \mu^m \underline{Z}^m \underline{e}_b \ ,$$

$$= \mu^m \quad E_b \underline{x}_1 + \sum_{k=2}^{n} \lambda_k^m \ \underline{x}_k \ \underline{y}_k' \ \underline{e}_b \quad . \qquad (10.24)$$

The theoretical increment associated with the steady state model in equation (10.13) can also be written in these terms

$$\tilde{\underline{e}}_{nb}(m) = \mu^m \ \tilde{\underline{Z}} \ \underline{e}_b \ ,$$

$$= \mu^m E_b \ \underline{x}_1 \ . \qquad (10.25)$$

The measures to be introduced here relate these actual and theoretical increments and a comparison of equations (10.24) and (10.25) reveals that such measures will be based on the convergence term suggested

previously in equation (10.22).

The first measure is based on computing the relative displacement from the steady state in percentage terms. This involves a ratio of the absolute differences in non-basic employment from the steady state to total non-basic employment and is stated as

$$\theta(m) = 100 \sum_{\ell=1}^{m} \mu^{\ell} \left| \sum_{k=2}^{n} \lambda_k^{\ell} \underline{x}_k \underline{y}_k' \underline{e}_b \right| / \mu(1 - \mu)^{-1} E_b. \qquad (10.26)$$

The other two measures are based on more conventional vector norms and contain the influence of the scale of non-basic employment to be allocated. A conventional distance norm reflecting differences between the actual and steady states can be calculated. Then

$$\begin{aligned} \emptyset(m) &= |\underline{e}_{nb}(m) - \tilde{\underline{e}}_{nb}(m)| / |\tilde{\underline{Z}}\, \underline{e}_b| \\ &= \{[\underline{e}_{nb}(m) - \tilde{\underline{e}}_{nb}(m)]'[\underline{e}_{nb}(m) - \tilde{\underline{e}}_{nb}(m)] / [\underline{e}_b' \tilde{\underline{Z}}' \tilde{\underline{Z}}\, \underline{e}_b]\}^{\frac{1}{2}}. \qquad (10.27) \end{aligned}$$

Using the definitions in equations (10.24) and (10.25), $\emptyset(m)$ can be made explicit in the following sense

$$\emptyset(m) = \mu^m \{[\underline{e}_b' \left( \sum_{k=2}^{n} \underline{y}_k \underline{x}_k' \lambda_k^m \right) \left( \sum_{k=2}^{n} \lambda_k^m \underline{x}_k \underline{y}_k' \right) \underline{e}_b] / [E_b^2 \underline{x}_1' \underline{x}_1]\}^{\frac{1}{2}}. \qquad (10.28)$$

Note that $\emptyset(m)$ depends on $\mu^m$ in a fairly simple way, and that when the actual distribution is the steady state distribution, $\emptyset(m) = 0$; the theoretical upper bound for $\emptyset(m)$ is $\mu^m$. This norm might constitute a good basis for comparisons between different applications if, for example, $\emptyset(1)$ were to be used. The last norm to be introduced has somewhat different properties to $\emptyset(m)$ and in some applications, this might be preferable. Then

$$\Omega(m) = \{[\underline{e}'_{nb}(m)\ \tilde{\underline{e}}_{nb}(m)]^{\frac{1}{2}}/|\tilde{\underline{Z}}\ \underline{e}_b|\}$$

$$= \{[\underline{e}'_{nb}(m)\ \tilde{\underline{e}}_{nb}(m)]/[\underline{e}'_b\ \tilde{\underline{Z}}'\tilde{\underline{Z}}\ \underline{e}_b]\}^{\frac{1}{2}} \qquad (10.29)$$

Equation (10.29) can be made more explicit using the spectral decomposition forms for the relevant vectors and thus

$$\Omega(m) = \mu^m \{[\underline{e}'_b(\sum_{k=1}^{n} \underline{y}_k\ \underline{x}'_k\ \lambda_k^m)\ \underline{x}_1 E_b]/[E_b^2 \underline{x}'_1\ \underline{x}_1]\}^{\frac{1}{2}}, \qquad (10.30)$$

from which it is clear that $\Omega(m)$ varies around $\mu^m$. When the predicted distribution $\underline{e}_{nb}(m)$ is equal to its steady state distribution, $\Omega(m)$ is equal to $\mu^m$. A less desirable feature of this norm is that if $\tilde{\underline{e}}_{nb}$ (m) is uniform, $\Omega(m)$ is also equal to $\mu^m$ regardless of $\underline{e}_{nb}(m)$. For each of these three measures, the eigenstructure of $\underline{Z}$ need not be extracted by computation for the convergence term can be computed directly from the difference between actual and steady state predictions.

EMPIRICAL DEMONSTRATIONS.

Schinnar's (1978) 3 zone example which he took from Roger's (1971) exposition of the Garin-Lowry model, has been reworked using the analytical techniques for measuring invariance proposed here. The emphasis in this chapter is on extracting the eigenstructure of the distribution matrix $\underline{Z}$ and on evaluating the proportion of invariance in the predicted distribution of non-basic employment, and the speed of convergence to the steady state. The data and computation associated with the model are presented in Table 10.1; the original matrix $\underline{Z}$, the theoretical steady state matrix $\tilde{\underline{Z}}$ and the associated inverses $(\underline{I} - \mu\ \underline{Z})^{-1}$ and $(\underline{I} - \mu\ \tilde{\underline{Z}})^{-1}$ are also shown

Table 10.1. Spatially Variant and Invariant Distributions Associated with the Schinnar-Rogers Example of the Garin-Lowry Model.

$$\underline{\frac{P}{Q}} = \begin{bmatrix} 0.5000 & 0.2500 & 0.2500 \\ 0.2500 & 0.500 & 0.2500 \\ 0.2500 & 0.2500 & 0.5000 \end{bmatrix} \qquad \underline{e}_b = \begin{bmatrix} 48 \\ 48 \\ 48 \end{bmatrix} \begin{matrix} \alpha^{-1} = 3{,}000 \\ \beta = 0.1666 \\ \mu = 0.5000 \end{matrix}$$

$$\underline{Z} = \begin{bmatrix} 0.3750 & 0.3125 & 0.3125 \\ 0.3125 & 0.3750 & 0.3125 \\ 0.3125 & 0.3125 & 0.3750 \end{bmatrix} (\underline{I} - \mu\underline{Z})^{-1} = \begin{bmatrix} 1.3548 & 0.3226 & 0.3226 \\ 0.3226 & 1.3548 & 0.3226 \\ 0.3226 & 0.3226 & 1.3548 \end{bmatrix}$$

$$\underline{\tilde{Z}} = \begin{bmatrix} 0.3333 & 0.3333 & 0.3333 \\ 0.3333 & 0.3333 & 0.3333 \\ 0.3333 & 0.3333 & 0.3333 \end{bmatrix} (\underline{I} - \mu\underline{\tilde{Z}})^{-1} = \begin{bmatrix} 1.3333 & 0.3333 & 0.3333 \\ 0.3333 & 1.3333 & 0.3333 \\ 0.3333 & 0.3333 & 1.3333 \end{bmatrix}$$

$$\underline{e}_{nb} = \{(\underline{I} - \mu\underline{Z})^{-1} - \underline{I}\}\underline{e}_b = \{(\underline{I} - \mu\underline{\tilde{Z}})^{-1} - \underline{I}\}\underline{e}_b = \begin{bmatrix} 48 \\ 48 \\ 48 \end{bmatrix}$$

in Table 10.1 and a casual comparison of $\underline{Z}$ and $\tilde{\underline{Z}}$ immediately suggests that $\underline{Z}$ is close to its steady state. Accordingly, the spectral decomposition of $\underline{Z}$ is presented in Table 10.2. There is a particular advantage to using the Rogers' (1971) example for this purpose. Because $\underline{Z}$ is symmetric, it is known that all its eigenvalues are real and thus the analysis is not complicated by having to deal with imaginary parts. Furthermore, the eigenvalues of $\underline{Z}$ need not be distinct to achieve a diagonalisation of $\underline{Z}$ into a set of orthogonal eigenvectors. Because $\underline{Z}$ is symmetric, the diagonalisation into $\underline{Z} = \underline{X}\ \underline{\Lambda}\ \underline{X}^{-1} = (\underline{Y}')^{-1}\underline{\Lambda}\ \underline{Y}'$ is simplified due to the fact that $\underline{X}^{-1} = \underline{X}'$ and $(\underline{Y}')^{-1} = \underline{Y}$. Therefore the diagonalisation which has been used here is based on $\underline{Z} = \underline{X}\ \underline{\Lambda}\ \underline{X}'$.

In Table 10.2, $\underline{Z}$ is presented in terms of its eigenstructure $\underline{X}\ \underline{\Lambda}\ \underline{X}'$ and its spectral decomposition $\Sigma_{k=1}^{n} \lambda_k\ \underline{x}_k\ \underline{x}_k'$. It is easily checked that $\underline{X}\ \underline{X}' = \underline{I}$ and thus the eigenvectors $\underline{x}_k$, k = 1, 2, ..., n constitute an orthonormal basis for the spectral decomposition. In fact, $\underline{Z}$ does not have 3 distinct eigenvalues: $\lambda_1$ is the dominant root and $\lambda_2 = \lambda_3$. In this example, this multiplicity of roots does not cause any difficulties in diagonalisation due to the fact that $\underline{Z}$ is symmetric. In any example where $\underline{Z}$ was asymmetric and multiple roots occurred, a small perturbation of $\underline{Z}$ should be sufficient to effect the diagonalisation. In this particular example, the eigenvectors $\underline{x}_k$, k = 1, 2,..., n were computed using the Householder method for symmetric matrices outlined in Fox (1964). Note that each eigenvector is normalised to $|\underline{x}_k| = 1$. There is however a limitation which arises from dealing with symmetric matrices which are stochastic. To achieve the required symmetry, $\underline{Z}$ must be doubly stochastic and

Table 10.2. Diagonalisation and Spectral Decomposition of the Matrix $\underline{Z}$

$$\underline{Z} = \underline{X}\,\underline{\Lambda}\,\underline{X}' = \begin{bmatrix} -0.5773 & 0.3869 & 0.7190 \\ -0.5773 & -0.8161 & -0.0245 \\ -0.5773 & 0.4292 & -0.6945 \end{bmatrix} \begin{bmatrix} 1.0000 & & \\ & 0.0625 & \\ & & 0.0625 \end{bmatrix} \begin{bmatrix} -0.5773 & -0.5773 & -0.5773 \\ 0.3869 & -0.8161 & 0.4292 \\ 0.7190 & -0.0245 & -0.6945 \end{bmatrix}$$

$$\underline{Z} = \sum_{k=1}^{3} \lambda_k \underline{x}_k \underline{x}_k' = 1.0 \begin{bmatrix} 0.3333 & 0.3333 & 0.3333 \\ 0.3333 & 0.3333 & 0.3333 \\ 0.3333 & 0.3333 & 0.3333 \end{bmatrix} + 0.0625 \begin{bmatrix} 0.1497 & -0.3157 & 0.1661 \\ -0.3157 & 0.6661 & -0.3503 \\ 0.1661 & -0.3503 & 0.1843 \end{bmatrix} + 0.0625 \begin{bmatrix} 0.5170 & -0.0176 & -0.4994 \\ -0.0176 & 0.0006 & 0.0170 \\ -0.4994 & 0.0170 & 0.4824 \end{bmatrix}$$

$$\underline{e}_{nb} = \{(\underline{I} - \mu\underline{Z})^{-1} - \underline{I}\}\hat{\underline{e}}_b = \begin{bmatrix} 51.0968 \\ 46.4516 \\ 46.4516 \end{bmatrix} \qquad \underline{e}_{nb} = \{(\underline{I} - \mu\tilde{\underline{Z}})^{-1} - \underline{I}\}\,\hat{\underline{e}}_b = \begin{bmatrix} 48 \\ 48 \\ 48 \end{bmatrix}$$

thus each zone in the model has the same distributional pattern in relation to every other zone. In this example, the initial distribution of basic employment is also uniform and thus *any* doubly stochastic matrix used to redistribute a uniform distribution will lead to the same uniform distribution. In this case, as the original matrix $\underline{Z}$ is doubly stochastic, so must be the steady state matrix $\tilde{\underline{Z}}$ and the associated distributions of non-basic employment are identical. Formally $\underline{e} = (\underline{I} - \mu\, \underline{Z})^{-1}\underline{e}_b = (\underline{I} - \mu\, \tilde{\underline{Z}})^{-1}\underline{e}_b$. This in fact is an excellent example of equifinality in urban distribution processes and the ability of the formal analysis to identify and handle such conditions represents an interesting line of research emerging from this chapter. However in this context, the implications of this equifinality are more immediate. Because the evaluation of invariance is to be based on statistics which measure differences in the location not interaction patterns, these statistics break down due to the fact that $\underline{e}_{nb}(m) = \tilde{\underline{e}}_{nb}(m)$, $\forall$ m. In short, a full analysis of distributional invariance which copes with equifinality must be based on extracting the complete eigen-structure as in Table 10.2 whereas the statistics in equations (10.26), (10.28) and (10.30) only reveal invariance in location. Thus to test these statistics, a non-uniform distribution of basic employment has been chosen, that is, $\hat{\underline{e}}_b' = [144 \quad 0 \quad 0]$ and the predictions from the model in terms of non-basic employment are contained in Table 10.2.

Table 10.3 presents the convergence of the model to its steady state in terms of the eigenvalues, $\lambda_1$, $\lambda_2 = \lambda_3$, and the three statistics $\theta(m)$, $\emptyset(m)$ and $\Omega(m)$. The previous casual observation that $\underline{Z}$ is extremely close to $\tilde{\underline{Z}}$ is confirmed from these results in that $\theta(m > 5)$ is about

Table 10.3. Convergence to the Steady State (Invariant) Distribution.

| Number of Iterations | Eigenvalues | | Convergence Measures | | |
|---|---|---|---|---|---|
| $m$ | $\lambda_1^m$ | $\lambda_2^m = \lambda_3^m$ | $\theta(m)$ | $\emptyset(m)$ | $\Omega(m)^{\dagger}$ |
| 1 | $10^0$ | $10^{-1.2041}$ | 4.1666 | $10^{-1.3546}$ | 0.5000 |
| 2 | $10^0$ | $10^{-2.4082}$ | 4.2969 | $10^{-2.8598}$ | 0.2500 |
| 3 | $10^0$ | $10^{-3.6124}$ | 4.3009 | $10^{-4.3649}$ | 0.1250 |
| 4 | $10^0$ | $10^{-4.8165}$ | 4.3011 | $10^{-5.8701}$ | 0.0625 |
| 5 | $10^0$ | $10^{-6.0206}$ | 4.3011 | $10^{-7.3752}$ | 0.0312 |

† Note that $\Omega(m) = \mu^m$ which is due to the fact that $\tilde{\underline{e}}_{nb}(m)$ is uniform. This in turn results from the doubly stochastic nature of $\underline{Z}$ and $\tilde{\underline{Z}}$.

4.3 percent, thus indicating that the final distribution of non-basic employment only differs from the theoretical invariant distribution by this order. In this case, as $\underline{Z}$ is doubly stochastic and symmetric, the original highly non-uniform distribution leads to a predicted distribution which is only 4.3 percent different from uniformity. Another conclusion is striking and this relates to the speed of convergence. Markov processes are well-known for their fast convergence and in this case, the process of distribution is to all intents and purposes in the steady state by the third iteration. The example is perhaps somewhat artificial in that the distribution matrices $\underline{Q}$ and $\underline{P}$ which form $\underline{Z}$ imply a higher level of interaction than is typical for British cities, although the conclusion that in many real applications, high levels of locational invariance with respect to basic employment exist, is inescapable.

The essential point of this analysis however is to establish the importance of the invariance property in urban modelling. It has been shown here that such invariance is a natural and necessary consequence of sequential distribution using stationary transition matrices, and given the fast convergence of the Markov distribution process, this property is highly significant. Clearly, the absolute importance of the property is affected by the amount of activity to be distributed, by $\mu$, and thus there are a whole range of different cases still to be analysed. Indeed, if invariance is as significant as in the example analysed here, there is a strong case for abandoning the spatial economic base framework and devising a different procedure, as implied by Schinnar (1978). Yet the Garin-Lowry model is still entirely appropriate if it is interpreted as an accounting framework (Batty, 1978) and the new form for the model given previously in

equation (10.23) demonstrates that the structure of the model is separable into spatially dependent and spatially independent components relating to the initial distribution of basic employment. This is perhaps the most important result of this chapter. One final point relating to invariance is worth making before conclusions are stated. The invariance in $\underline{Z}$ can arise from invariance in $\underline{P}$ or $\underline{Q}$ or both, or in the interaction between them. If $\underline{P}$ or $\underline{Q}$ is idempotent, $\underline{Z}$ will be idempotent. There are many types of situation to explore with regard to the source of invariance and this represents another important line of further research.

CONCLUSIONS.

Following Schinnar's (1978) heuristic demonstration of invariant distributional regularities in models of the Garin-Lowry type, this chapter has sought to extend some of his analysis in more formal terms and link it to earlier ideas concerning pseudo-dynamic models. It has been shown that all models of this type are subject to distributional invariance, that their distribution processes display the Markov property. A natural consequence of this analysis is a new form for such models which is based on a decomposition into components associated with different determinants of spatial variation. But there are many implications arising from this analysis yet to be explored and these form a catalogue of items for research. An important line of applied research must focus upon real examples and the degree of invariance associated with each one. Furthermore, it may be possible to show that such invariance is related to the level of income/wealth in particular cities and to thus begin to reassess the economic base logic in this light. Some further

theoretical analysis of the ways in which different degrees of invariance can arise in such models is warranted, and the somewhat baffling problem of equifinality is worthy of extensive investigation.

The model dealt with here is the simplest in a wider family of such models and it is possible to extend the analysis to more general models as is shown in the next chapter. In particular, the effect of locational constraints on invariance, the uniqueness of any particular degree of invariance, and the effect of the spatial system on the degree of invariance are all important questions. The causal structure of this type of model is also revealed by this inquiry into invariance. The decomposition of the model into dependent and independent components opens up the perennial problem of causality in these types of model in terms of the influence of basic on non-basic employment and vice versa. Moreover, the way in which variables are combined in linear additive fashion is also a subject for concern; for example, the fact that one set of distribution rules if idempotent, is able to dominate another set is a disturbing feature and suggests areas in which such models might be extended and improved. This list of suggestions constitutes a large task but all the tools for further analysis have been introduced in this chapter. In subsequent applications, it would appear essential to identify the degree of distributional invariance if only as an aid to interpretation but hopefully as a prelude to the design of more appropriate models. Some suggestions in this regard will be explored in the next and last substantive chapter of this thesis.

# CHAPTER 11.

## LINEAR ANALYSIS OF URBAN MODELS.

In this chapter, the change in emphasis towards more formal analysis of the dynamics of urban model structures will be continued in a number of ways. The analysis introduced in the last chapter will be deepened and generalised, and then applied empirically to urban spatial structure in Melbourne. The conventional urban model of the Lowry type will itself be generalised, not as comprehensively as in Chapter 2 but sufficient to enable the Markov analysis to be broadened. Moreover, in this chapter some sense of closure to this thesis will be implicit in that some of the generalisations contained in the review in Chapter 2 will be presented, and the argument linked to those distinctions between linearity and nonlinearity which serve to classify recent work in urban modelling.

In this latter regard as developed in Chapter 2, it is clear that urban models are frequently characterised as being predominantly structured in linear or nonlinear terms, but in several contemporary developments, both linear and nonlinear styles of analysis are intermixed. In spatial interaction modelling for example, model derivation is usually through nonlinear optimisation leading to nonlinear model

forms as seen here in the use of information-minimising while when such models are coupled together to form more general structures, such coupling is usually effected through linear accounting, subject to linear constraints. Thus these models can be analysed using both linear and nonlinear analysis, each type of analysis emphasising different properties of the model structure.

The traditional Lowry model is the classic example. The model was first stated by Lowry (1964) as an implicitly nonlinear structure. It was then developed in linear terms by Garin (1966) and Harris (1966) using analogies with input-output models, and then in nonlinear terms by Wilson (1974) who emphasised the derivation of its spatial interaction components through entropy-maximising. More recently, attention has been directed at coupling and solving the model's spatial interaction components in a more general nonlinear optimisation framework in which the model's linear structure is implicitly represented through its constraints (Wilson, Coelho, Macgill and Williams, 1981). These developments were presented in Chapter 2.

Most recent work has, in fact, been directed to the nonlinear analysis of such models, and it is perhaps surprising that so little work has been concerned with exploring the model's linear structure, especially as such models appear more structurally transparent, and easier to extend and adapt in linear terms. Moreover, linear models have been developed extensively in regional science but hitherto, there have been few attempts to generalise this class of models to explore common properties. It is the purpose of this chapter to present such a generalisation and within this, to begin to explore the effect of

model structure on performance in terms of the balance of input and output variables, as well as clarifying the question of choice of an appropriate model structure. To this end, these ideas will be illustrated using those linear urban models forming the subject of this thesis in which activities are spatially distributed although these ideas are also applicable to input-output, social exchange and various demographic accounts-based models.

The general model structure pertaining to this class is first stated and then adapted to two activities, the spatial distributions of population and employment. Reduced forms are derived and various model types dependent upon different spatial distributional assumptions and input variables are characterised, including the Lowry model (Lowry, 1964; Batty, 1976) and Coleman's (1973) model of collective action based on the theory of social exchange. The effect of different distributional assumptions is then explored using the eigenstructure analysis developed in Chapter 10 and this gives a fairly comprehensive picture of the degree to which spatial solutions to these models are determined by model structure, input data and particular transformations.

Applications to an eight zone model of Melbourne then serve to give these findings some empirical credibility. This suggests that much more research is required into invariance characterising spatial model solutions, the choice of inputs and outputs in particular model applications, and the level of resolution appropriate to any application. Although these issues pertain to the linear urban models discussed here, they are of more general import, and by way of conclusion, certain rules of thumb for good model design are presented. The

notation used here differs from that in Chapter 10 in that it is consistent with earlier chapters and particularly Chapter 2. The main results of Chapter 10 will thus be repeated in the conventional notation, thus enabling the reader to treat the generalisations introduced here in a self-contained manner.

LINEAR STRUCTURES AND SOLUTIONS FOR URBAN MODELS.

To introduce the framework, first consider an activity $y_1$ distributed over n zones or sectors, and two activities $y_2$ and $x_2$ distributed over m zones or sectors. If $\underline{y}_2$ and $\underline{x}_2$ are 1xm row vectors, $\underline{y}_1$ is a 1xn row vector and $\underline{A}_1$ is an nxm matrix which transforms $\underline{y}_1$ into $\underline{y}_2$, then the linear model can be written as

$$\underline{y}_2 = \underline{y}_1\underline{A}_1 + \underline{x}_2 \quad . \tag{11.1}$$

In a similar manner to equation (11.1), it is possible to develop a chain of relationships in which $\underline{y}_z$ can be predicted as the sum of a transformation of $\underline{y}_{z-1}$, and exogenous variables $\underline{x}_z$. For all variables $\underline{y}_z$ to be predicted however, the chain must be closed; that is at some point, $\underline{y}_z = \underline{y}_1$. The simplest possible case from equation (11.1) is where $\underline{y}_2 = \underline{y}_1$ which would imply m=n, and in this case, equation (11.1) could represent the structure of a conventional input-output model. In this context, it is necessary to examine the more general case where $z \geqslant 2$, and thus a suitable example of the closed sequence involves another equation for $\underline{y}_1$ given as

$$\underline{y}_1 = \underline{y}_2\underline{A}_2 + \underline{x}_1 \quad , \tag{11.2}$$

where $\underline{x}_1$ is a 1xn row vector and $\underline{A}_2$ an mxn transformation matrix.

Solutions to the system of equations in (11.1) and (11.2) are given by the following reduced forms which result from substituting equations (11.1) into (11.2) and (11.2) into (11.1). These are

$$\underline{Y}_1 = \underline{y}_1\underline{A}_1\underline{A}_2 + \underline{x}_2\underline{A}_2 + \underline{x}_1 \quad , \qquad \text{and} \qquad (11.3)$$

$$\underline{y}_2 = \underline{y}_2\underline{A}_2\underline{A}_1 + \underline{x}_1\underline{A}_1 + \underline{x}_2 \quad . \qquad (11.4)$$

In one sense, equations (11.3) and (11.4) might be considered duals of one another. Explicit solutions for $\underline{y}_1$ and $\underline{y}_2$ can be given by rearranging (11.3) and (11.4), but at this stage, it is more appropriate to consider their solution in matrix split or iterative form as

$$\underline{y}_1(t) = \underline{y}_1(0)(\underline{A}_1\underline{A}_2)^t + (\underline{X}_2\underline{A}_2 + \underline{x}_1) \sum_{\tau=0}^{t-1} (\underline{A}_1\underline{A}_2)^\tau, \quad \text{and} \qquad (11.5)$$

$$\underline{y}_2(t) = \underline{y}_2(0)(\underline{A}_2\underline{A}_1)^t + (\underline{x}_1\underline{A}_1 + \underline{x}_2) \sum_{\tau=0}^{t-1} (\underline{A}_2A_1)^\tau. \qquad (11.6)$$

$\underline{y}_1(0)$ and $\underline{y}_2(0)$, $\underline{y}_1(t)$ and $\underline{y}_2(t)$ represent the starting, and t'th solution vectors to $\underline{y}_1$ and $\underline{y}_2$ respectively, while $(\underline{A}_1\underline{A}_2)^0$ and $(\underline{A}_2\underline{A}_1)^0$ are nxn, and mxm identity matrices. A general solution for the case where z = N, N > 2 has already been given in Chapter 2.

Whether or not equations (11.5) and (11.6) converge to unique vectors $\underline{y}_1$ and $\underline{y}_2$ will depend upon the properties of $\underline{A}_1$ and $\underline{A}_2$. In this form, these iterative solutions refer to the static equilibrium framework based on equations (11.1) to (11.4). However equations (11.5) and (11.6) could refer to dynamic versions of (11.3) and (11.4) with fixed inputs, and thus the framework developed here could easily be extended to cover dynamic models, for example the manpower and

educational planning models of the type discussed by Bartholomew (1982). Furthermore if there are no inputs to these models, that is if $\underline{x}_1 = \underline{0}$ and $\underline{x}_2 = \underline{0}$, equations (11.5) and (11.6) are similar to those of a first order process which if $\underline{A}_1$ and $\underline{A}_2$ were stochastic matrices would be a Markov process, equivalent to that developed by Coleman (1973). These possibilities will be explored further in the sequel.

To develop this framework further, it is necessary to make specific assumptions about the types of distribution and transformation involved which follow the ideas of previous chapters. Then two urban activities, employment $e_i$, $i = 1,2,\ldots,n$ and population $p_j$, $j = 1,2,\ldots,m$ are defined in spatial distributional terms so that

$$\sum_i e_i = 1 \quad \text{and} \quad \sum_j p_j = 1 \quad ,$$

and these are related through

$$e_i = \beta a_i + (1-\beta) b_i \quad , \quad 0 \leqslant \beta \leqslant 1. \tag{11.7}$$

$a_i$ is service and $b_i$ basic employment in i, normalised so that

$$\sum_i a_i = \sum_i b_i = 1 \quad .$$

$\beta$ is the ratio of service to total employment in the system.

Population is also considered as the sum of two components, internal population $g_j$ and external (or basic) population $h_j$ which in distributional terms are defined so that

$$\sum_j g_j = \sum_j h_j = 1 \quad .$$

Population is then given by

$$p_j = \psi g_j + (1-\psi)h_j \ , \qquad 0 \leqslant \psi \leqslant 1 \ , \tag{11.8}$$

where $\psi$ is the ratio of internal to total population.

It is assumed that basic employment and external population are exogenous variables equivalent to $\underline{x}_1$ and $\underline{x}_2$ defined earlier, and that service employment and internal population are endogenous; service employment is modelled as a linear function of population, and internal population as a linear function of employment defined respectively as

$$a_k = \sum_j p_j B_{jk} \ , \qquad \sum_k B_{jk} = 1 \ , \qquad \text{and} \tag{11.9}$$

$$g_j = \sum_i e_i A_{ij} \ , \qquad \sum_j A_{ij} = 1 \ . \tag{11.10}$$

$B_{jk}$ and $A_{ij}$ are stochastic matrices which measure the demand by the population in zone j for services in zone k, and the demand by employees in zone i for housing in zone j respectively. These transformations are consistent with well-established ideas concerning spatial interaction (Wilson, 1974) and it is assumed that each spatial transformation matrix is strongly-connected in the graph or network-theoretic sense.

Equations (11.7) and (11.8) can now be written in linked form. Substituting for $a_i$ in (11.7) from (11.9), and $g_j$ in (11.8) from (11.10) leads to

$$e_k = \beta \sum_j p_j B_{jk} + (1-\beta)b_k \ , \qquad \text{and}$$

$$p_j = \psi \sum_i e_i A_{ij} + (1-\psi)h_j \ .$$

In matrix terms, these equations can be written as

$$\underline{e} = \beta \underline{p}\ \underline{B} + (1-\beta)\ \underline{b} \quad , \qquad \text{and} \qquad (11.11)$$

$$\underline{p} = \psi \underline{e}\ \underline{A} + (1-\psi)\ \underline{h} \quad . \qquad (11.12)$$

Comparing equations (11.11) with (11.2), and (11.12) with (11.1), it is clear that $\underline{e} = \underline{y}_1$, $\underline{p} = \underline{y}_2$, $\beta\underline{B} = \underline{A}_2$, $\psi\underline{A} = \underline{A}_1$, $(1-\beta)\underline{b} = \underline{x}_1$, and $(1-\psi)\underline{h} = \underline{x}_2$, thus the reduced forms in (11.3) and (11.4) and the matrix iterative solutions in (11.5) and (11.6) apply. It is however possible to say more about solutions to this model for the properties of $\underline{A}_1$ and $\underline{A}_2$ have now been specified.

In analogy to equation (11.3), the reduced form for employment in equation (11.11) is given as

$$\underline{e} = \psi\beta \underline{e}\ \underline{A}\ \underline{B} + \beta(1-\psi)\ \underline{h}\ \underline{B} + (1-\beta)\underline{b} \quad . \qquad (11.13)$$

The three terms on the RHS of (11.13) reflect service employment generated indirectly from basic employment and external population ($\psi\beta \underline{e}\ \underline{A}\ \underline{B}$), service employment generated directly from external population ($\beta(1-\psi)\underline{h}\ \underline{B}$), and basic employment ( $(1-\beta)\ \underline{b}$). The ratios $\psi\beta$, $\beta(1-\psi)$ and $(1-\beta)$ also reflect the fractions of such employment in the model solution. In similar fashion, the reduced form of equation (11.12) is given as

$$\underline{p} = \psi\beta\ \underline{p}\ \underline{B}\ \underline{A} + \psi(1-b)\ \underline{b}\ \underline{A} + (1-\psi)\underline{h} \quad , \qquad (11.14)$$

where $\psi\beta\underline{p}\ \underline{B}\ \underline{A}$ is the component of population associated with service employment, $\psi(1-\beta)\underline{b}\ \underline{A}$ is the population directly associated with basic employment, and $(1-\psi)\underline{h}$ is external population. Note that the ratios $\psi\beta$, $\psi(1-\beta)$ and $(1-\psi)$ reflect the fractions of each of these components in the final solution.

The matrix iterative solutions of equations (11.13) and (11.14) starting from arbitrary but normalised vectors $\underline{e}(0)$ and $\underline{p}(0)$ can be stated in analogy to equations (11.5) and (11.6) as

$$\underline{e}(t) = (\psi\beta)^t \underline{e}(0)(\underline{A}\ \underline{B})^t + [\beta(1-\psi)\underline{h}\ \underline{B} + (1-\beta)\underline{b}] \sum_{\tau=0}^{t-1} (\psi\beta)^\tau (\underline{A}\ \underline{B})^\tau,$$

and (11.15)

$$p(t) = (\psi\beta)^t \underline{p}(0)(\underline{B}\ \underline{A})^t + [\psi(1-\beta)\underline{b}\ \underline{A} + (1-\psi)\underline{h}] \sum_{\tau=0}^{t-1} (\psi\beta)^\tau (\underline{B}\ \underline{A})^\tau. \quad (11.16)$$

First we will assume that $0 < \psi\beta < 1$, in short that exogenous inputs exist and endogenous activity is generated from it, and then in the next section we will explore the special case where all activities are endogenously determined, $\psi\beta = 1$. Thus for $0 < \psi\beta < 1$, $(\psi\beta)^t \to 0$ as $t \to \infty$, and thus the first terms on the RHS of equations (11.15) and (11.16) converge to zero; the solutions thus depend on the second terms involving the matrix summations. As $\underline{A}\ \underline{B}$ and $\underline{B}\ \underline{A}$ are stochastic matrices, powers of these matrices will also be stochastic, so any convergence of these summations will depend on $\psi\beta$.

We remarked earlier that equations such as (11.15) and (11.16) can be regarded as duals of one another, thus it is only necessary to illustrate results for one of these, for the other can be determined from the equilibrium relations in equations (11.11) or (11.12). In the rest of this chapter,we will only consider solutions to the population equation, equation (11.16). Then in the limit,

$$\underline{p} = \lim_{t \to \infty} \underline{p}(t) = \lim_{t \to \infty} [\psi(1-\beta)\underline{b}\ \underline{A} + (1-\psi)\underline{h}] \sum_{\tau=0}^{t-1} (\psi\beta)^\tau (\underline{B}\ \underline{A})^\tau, \quad (11.17)$$

and as $(\psi\beta)^\tau (\underline{B}\ \underline{A})^t \to \underline{0}$ as $t \to \infty$, the summation term in (11.17) is a

converging geometric matrix series. It is easily shown that

$$\underline{p} = \lim_{t \to \infty} \underline{p}(t) = [\psi\ (1-\beta)\underline{b}\ \underline{A} + (1-\psi)\underline{h}][\underline{I} - \psi\beta\ \underline{B}\ \underline{A}]^{-1} \qquad (11.18)$$

which of course could have been derived directly from equation (11.14). Using equation (11.18) in (11.11) leads to

$$\underline{e} = \beta[\psi(1-\beta)\underline{b}\ \underline{A} + (1-\psi)\underline{h}][\underline{I} - \psi\beta\ \underline{B}\ \underline{A}]^{-1}\ \underline{B} + (1-\beta)\underline{b}\ ,$$

and a dual form for $\underline{p}$ and $\underline{e}$ exists by taking $\underline{e}$ from the convergence of equation (11.15) and substituting this into equation (11.12).

LINEAR DYNAMICS: MARKOVIAN URBAN MODELS.

The results given above in equations (11.7) to (11.8) relate to the case where each activity is determined partly by some exogenous activity and partly as a transformation of another endogenous activity. In the case of this spatial urban model, basic employment might be considered as export orientated employment in the traditional economic sense, or as employment whose location it is impossible to model (Massey, 1973). External population might reflect the same - either population dependent economically on activity outside the region or that whose location it is not possible to simulate, such as population located by some public agency. It is however instructive to examine three other variants of this model which reflect different balances of exogenous activity, thus different model structures.

First there is the case where there is no external population, that is where $\psi = 1$. The resulting model is in effect one in which employment and population are now a function only of basic employment, and the model in this form is the conventional Lowry model as can easily be seen by making the appropriate simplifications to equations (11.11) to (11.18).

The model in fact is the Lowry model in its matrix form (Harris, 1966; Garin, 1966). Second, there is the case where there is no basic employment, that is where $\beta = 1$. In this case, external population is the driving force of the model and this, it might be argued, is more appropriate for a model of a British New Town situation, say, where the population is located in a planned fashion. This model is the basic population equivalent of the Lowry model, and already the advantages of this general framework are becoming apparent in enabling the logic of linear urban models such as the Lowry model to be extended to other types of basic spatial determinant. In the sequel, these two single exogenous input models together with the model based on both inputs will be developed, but another interesting case of much greater analytic value emerges from the model with no exogenous inputs, that is where all population and employment are determined endogenously, where $\psi\beta = 1$.

It is worth examining this case in more detail. Equation (11.16) now becomes

$$\underline{p}(t) = \underline{p}(0)\ (\underline{B}\ \underline{A})^t \qquad (11.19)$$

and the solution depends on the behaviour of $(\underline{B}\ \underline{A})^t$. $\underline{B}\ \underline{A}$ is a stochastic matrix and assuming again that $\underline{B}\ \underline{A}$ is strongly-connected which is an essential assumption of the spatial interaction transformation in any case, $(\underline{B}\ \underline{A})^t$ converges to an idempotent matrix $\underline{Z}$ in which each row is identical. Then

$$\tilde{\underline{p}} = \lim_{t \to \infty} \underline{p}(t) = \underline{p}(0)(\underline{B}\ \underline{A})^t = \underline{p}(0)\underline{Z} \qquad (11.20)$$

From equation (11.20) it is clear that $\tilde{\underline{p}}$, the steady state population distribution, is equivalent to each identical row of the steady state

matrix $\underline{Z}$. As $\underline{Z}$ is idempotent, multiplication of (11.20) by $(\underline{B}\ \underline{A})$ leads to

$$\underline{\tilde{p}} = \underline{\tilde{p}}\ \underline{B}\ \underline{A} \quad ,$$

which is equivalent to equation (11.14) with $\psi\beta = 1$. Clearly the steady state population distribution is the steady state of a discrete Markov process as was last demonstrated in the last chapter. In similar fashion, it can be shown that $(\underline{A}\ \underline{B})^t$ converges to an idempotent matrix $\underline{Q}$ as $t \to \infty$ and using the same logic as above, the steady state employment $\underline{\tilde{e}}$ is given as

$$\underline{\tilde{e}} = \underline{\tilde{e}}\ \underline{A}\ \underline{B} \quad .$$

Furthermore, when $n = m$, $\underline{Q} = \underline{Z}\ \underline{B}$ and $\underline{Z} = \underline{Q}\ \underline{A}$; in short, when $\psi\beta = 1$, equations (11.15) and (11.16) represent dual Markov processes.

In this form, the model is equivalent to Coleman's (1973) model of collective action in which the equilibrium can be interpreted as the outcome of an exchange process. In fact, an exchange interpretation could quite easily be developed for urban models such as those in the same spirit as that developed by Coleman, thus enabling insights into these types of models to be further enriched (Batty, 1981). Moreover the framework developed here shows how the Coleman model might also be seen as a special case of a more general model of collective action in which such action is seen as being determined by both endogenous and exogenous factors. However such interpretations are beyond the immediate concern of this chapter. Wider implications for this work were sketched in Chapter 2.

The model without exogenous inputs, referred to hereafter as the Coleman model, although interesting in its own right as a distinct model

structure in the framework, is also useful in that it highlights the fact that the transformation matrix $(\underline{B}\ \underline{A})^t$ converges towards the idempotent matrix $\underline{Z}$ as the number of iterations of the process of solution increases. As any stochastic row vector multiplied by this idempotent matrix gives a row of this matrix, this implies that in the case where the transformation matrix is or becomes idempotent, the exogenous vector then has no influence on the resulting solution. This is the condition of invariant distributional regularity identified by Schinnar (1978) and analysed in Chapter 10.

To demonstrate this idea, consider the case where the matrix $\underline{B}\ \underline{A}$ is already idempotent, that is

$$\underline{B}\ \underline{A} = (\underline{B}\ \underline{A})^t = \underline{Z},\ t > 0.$$

Then for the case where $0 < \psi\beta < 1$, the matrix series in equation (11.17) can be written as

$$\lim_{t \to \infty} \sum_{\tau=0}^{t-1} (\psi\beta)^\tau (\underline{B}\ \underline{A})^\tau = \underline{I} + \underline{Z} \sum_{\tau=1}^{\infty} (\psi\beta)^\tau ,$$

$$= \underline{I} + \psi\beta(1-\psi\beta)^{-1}\ \underline{Z} \quad . \qquad (11.21)$$

Using equation (11.21) in (11.17), the equilibrium population referred to as the population from the steady state model, now becomes

$$\hat{\underline{p}} = [\psi(1-\beta)\underline{b}\ \underline{A} + (1-\psi)\underline{h}][\underline{I} + \psi\beta\ (1-\psi\beta)^{-1}\underline{Z}] ,$$

which simplifies, using the fact that each row of $\underline{Z}$ is $\tilde{\underline{p}}$, to

$$\hat{\underline{p}} = \psi\beta\ \tilde{\underline{p}} + \psi(1-\beta)\underline{b}\ \underline{A} + (1-\psi)\ \underline{h} . \qquad (11.22)$$

Equation (11.22) shows that population $\hat{\underline{p}}$ is a function of the input data, and of the steady state $\tilde{\underline{p}}$ and this implies that the input has no influence on the endogenously generated population.

A similar result holds for employment. Substituting equation (11.22) into (11.11) gives

$$\hat{\underline{e}} = \psi\beta\, \underline{b}\, \underline{A}\, \underline{B} + \beta(1-\psi)\underline{h}\, \underline{B} + (1-\beta)\underline{b} \ .$$

Now as $\underline{B}\, \underline{A} = \underline{Z}$, and $\underline{A}\, \underline{B}\, \underline{A} = \underline{A}\, \underline{Z} = \underline{Z} = \underline{Q}\, \underline{A}$, then $\underline{A}\, \underline{B} = \underline{Q}$, and the steady state employment can now be written as

$$\hat{\underline{e}} = \psi\beta\tilde{\underline{e}} + \beta(1-\psi)\underline{h}\, \underline{B} + (1-\beta)\, \underline{b} \ , \qquad (11.23)$$

which has the same structure as equation (11.22). From equations (11.22) and (11.23), it is clear that the degree to which $\hat{\underline{p}}$ approaches $\tilde{\underline{p}}$ and $\hat{\underline{e}}$ approaches $\tilde{\underline{e}}$ depends on the ratio $\psi\beta$. Three types of model based on equations (11.22) and (11.23) where $0 < \psi < 1$, and $0 < \beta < 1$, where $\psi = 1$ and where $\beta = 1$ will be developed in the applications presented below.

There are several different ways in which the matrices $\underline{B}\, \underline{A}$ and $\underline{A}\, \underline{B}$ may be idempotent. It is clear that if $\underline{A}$ or $\underline{B}$ is idempotent, then either of their products is idempotent. Then if $\underline{A}$ is idempotent, that is if the probability of residing in any place is independent of the place of employment, equation (11.22) simplifies to

$$\hat{\underline{p}} = \psi\tilde{\underline{p}} + (1-\psi)\underline{h} \ ,$$

while if $\underline{B}$ is idempotent, that is if the probability of demanding services in any place is independent of the place where that demand is generated, equation (11.23) simplifies to

$$\hat{\underline{e}} = \beta\tilde{\underline{e}} + (1-\beta)\underline{b} \ .$$

If both $\underline{A}$ and $\underline{B}$ are idempotent, then both equations (11.22) and (11.23) simplify in the manner shown.

Another possible effect of idempotence which will be developed in the sequel involves the situation where

$$\underline{B}\ \underline{A} = (\underline{B}\ \underline{A})^t = \underline{I}\ , \quad t > 0\ .$$

In this situation, equation (11.17) simplifies to

$$\bar{p} = \frac{\psi(1-\beta)}{(1-\psi\beta)}\ \underline{b}\ \underline{A} + \frac{(1-\psi)}{(1-\psi\beta)}\ \underline{h}\ , \tag{11.24}$$

and using equation (11.24) in (11.11) gives

$$\bar{e} = \frac{\beta(1-\psi)}{(1-\psi\beta)}\ \underline{h}\ \underline{B} + \frac{(1-\beta)}{(1-\psi\beta)}\ \underline{b}\ . \tag{11.25}$$

In this case, the equilibrium population and employment distributions are simply proportional to the appropriately scaled fraction of each exogenous distribution of activities. Clearly this situation can arise in several ways. For example if both $\underline{B}$ and $\underline{A}$ are identity matrices where $m = n$ then this implies no spatial interaction in the system whatsoever; that is employees live and work in the same zone and demand their services there. The same situation can also arise if the patterns of spatial demand for housing are the inverse of those for services, that is where $\underline{A} = \underline{B}^{-1}$ and $\underline{B} = \underline{A}^{-1}$. In all these cases, $\underline{B}\ \underline{A} = \underline{I}$ and $\underline{A}\ \underline{B} = \underline{I}$, but these patterns and the resulting identical locational distributions can clearly arise under very different conditions of spatial interaction.

In the applications which follow in a later section, three types of model will be developed; the model based on actual interaction matrices $\underline{A}$ and $\underline{B}$ given in equations (11.13) and (11.14), the model based on the steady state interaction patterns derived from $\underline{Z} = \lim_{t \to \infty}(\underline{B}\ \underline{A})^t$ and $\underline{Q} = \lim_{t \to \infty}(\underline{A}\ \underline{B})^t$

in equations (11.22) and (11.23), and the model based on equations (11.24) and (11.25) in which it is assumed that this pattern is associated with no interaction or self cancelling interaction. For each of these models, three model structures will be tested; first where both inputs are present, where $0 < \psi < 1$ and $0 < \beta < 1$, second where basic employment is the only input, where $\psi = 1$, and third where external population is the only input, where $\beta = 1$. Finally the Markov model with no inputs, where $\psi\beta = 1$, Coleman's model, will be developed, thus giving 10 different models in all to be explored.

THE MEASUREMENT OF DISTRIBUTIONAL INVARIANCE.

In the previous section, we indicated that in the absence of input data, the ultimate distribution of activities in the model would depend on their steady state distribution matrices $\underline{Z}$ and $\underline{Q}$. In the case where there are exogenous inputs and where the transformation matrices are already in the steady state, the solutions can be derived as a weighted sum of the steady state and input distributions. This suggests that in the case of the general model, it is possible to measure formally the degree to which the distribution matrices $\underline{B}\ \underline{A}$ and $\underline{A}\ \underline{B}$ approach the steady state when solutions to the model are derived. This was the conclusion of the last chapter and there a variety of measures of invariance were proposed. These will be extended here but as a precursor to their application, the main results of Chapter 10 will be sketched again.

To demonstrate the relationship, we will use $\underline{B}\ \underline{A}$ and its steady state $\underline{Z} = \lim_{t \to \infty} (\underline{B}\ \underline{A})^t$. Assuming that the eigenvalues of $\underline{B}\ \underline{A}$ are all distinct, the matrix $\underline{B}\ \underline{A}$ can be represented as

$$\underline{B}\ \underline{A} = \underline{R}^T\ \underline{\Lambda}\ \underline{S} = \sum_{k=1}^{m} \lambda_j \underline{r}_j^T \underline{s}_j = \sum_{j=1}^{m} \lambda_j \underline{V}_j \quad , \tag{11.26}$$

where $\underline{R}$ is an mxm matrix of right-hand eigenvectors of $\underline{B}\ \underline{A}$, $[\underline{r}_j]$, $\underline{S}$ is an mxm matrix of left-hand eigenvectors $[\underline{s}_j]$, and $\underline{\Lambda}$ is an mxm diagonal matrix of the m eigenvalues of $\underline{B}\ \underline{A}$ where each eigenvalue $\lambda_j$ on the diagonal $\Lambda_{jj}$ is associated with the eigenvectors $\underline{r}_j$ and $\underline{s}_j$. $\underline{V}_j = \underline{r}_j^T\ \underline{s}_j$ and this matrix is defined as the spectral set. Assuming that the scales of $\underline{s}_j$ and $\underline{r}_j^T$ are chosen so that $\underline{s}_j \underline{r}_j^T = 1$, then $\underline{V}_j$ satisfies the following relations

$$\underline{V}_\ell \underline{V}_j = \underline{0}\ , \quad \ell \neq j\ ; \quad \underline{V}_\ell \underline{V}_j = \underline{V}_j\ , \ \ell = j, \text{ and } \sum_{j=1}^{m} \underline{V}_j = \underline{I}. \tag{11.27}$$

These results are taken from Bailey (1964). Note that T indicates the matrix transpose operation, in contrast to the prime ' used in the last chapter.

The decomposition defined in equations (11.26) and (11.27) enables the powers of $\underline{B}\ \underline{A}$ to be expressed in simple form as

$$(\underline{B}\ \underline{A})^t = \sum_{j=1}^{m} \lambda_j^t\ \underline{V}_j \quad . \tag{11.28}$$

From the Perron-Frobenius theorem (see Heal, Hughes and Tarling, 1974), a stochastic matrix such as $\underline{B}\ \underline{A}$ has a dominant eigenvalue equal to 1, and all other eigenvalues of the matrix have an absolute value less than 1. Assuming these values are distinct (slight perturbation of the values in $\underline{B}\ \underline{A}$ will normally ensure this within an acceptable error bound for $\underline{B}\ \underline{A}$), then it is possible to order the eigenvalues and eigenvectors of $\underline{B}\ \underline{A}$ so that $\lambda_1(=1)>|\lambda_2|>|\lambda_3|>\ldots>|\lambda_m|$. In the case of the dominant eigenvalue $\lambda_1=1$,

$$\lambda_1^t \underline{V}_1 = \lambda_1 \underline{r}_1^T \underline{s}_1 = \underline{1}^T \underline{s}_1 = \underline{Z} \quad , \tag{11.29}$$

because the right-hand eigenvector associated with $\lambda_1 = 1$ must be the unit vector, that is $\underline{B}\ \underline{A}\ \underline{1}^T = \underline{1}^T$ and $\underline{s}_1$ represents the steady state vector associated with $\underline{B}\ \underline{A}$, that is $\underline{s}_1 = \underline{s}_1\ \underline{B}\ \underline{A}$, the left-hand eigenvector. Using the equation (11.29) in (11.28), it is possible to write the powers of $\underline{B}\ \underline{A}$ as

$$(\underline{B}\ \underline{A})^t = \underline{Z} + \sum_{j=2}^{m} \lambda_j^t\ \underline{V}_j \quad , \tag{11.30}$$

where it is clear that in the limit as $t \to \infty$, $(\underline{B}\ \underline{A})^t \to \underline{Z}$, and $\sum_{j=2}^{m} \lambda_j^t\ \underline{V}_j \to \underline{0}$. Thus the difference between the matrix $\underline{B}\ \underline{A}$ and its steady state has the simple form

$$\underline{B}\ \underline{A} - \underline{Z} = \sum_{j=2}^{m} \lambda_j\ \underline{V}_j \quad ,$$

and this difference converges to $\underline{0}$ as $t \to \infty$.

It is now possible to represent the matrix series which arises in the iterative solution to the model in equation (11.17) using (11.28) as

$$\sum_{\tau=0}^{\infty} (\psi\beta)^\tau\ (\underline{B}\ \underline{A})^\tau = \sum_{\tau=0}^{\infty} \sum_{j=1}^{m} (\psi\beta\lambda_j)^\tau\ \underline{V}_j . \tag{11.31}$$

Then as $|\lambda_j| \leq 1$ and $0 < \psi\beta < 1$, the series in (11.31) can be simplified as follows

$$\sum_{\tau=1}^{\infty} (\psi\beta\lambda_j)^\tau = (1 - \psi\beta\lambda)^{-1} \quad , \qquad \text{and}$$

$$\sum_{\tau=1}^{\infty} (\psi\beta\lambda_j)^\tau = \psi\beta\lambda_j \sum_{\tau=0}^{\infty} (\psi\beta\lambda_j)^\tau = \psi\beta\lambda_j(1-\psi\beta\lambda_j)^{-1} .$$

Equation (11.31) can now be written as

$$\sum_{\tau=0}^{\infty} (\psi\beta)^{\tau}(\underline{B}\ \underline{A})^{\tau} = \sum_{j=1}^{m} (1-\psi\beta\lambda_j)^{-1}\underline{V}_j \quad ,$$

$$= \underline{I} + \sum_{j=1}^{m} \psi\beta\lambda_j(1-\psi\beta\lambda_j)^{-1}\underline{V}_j \ ,$$

$$= \underline{I} + \psi\beta(1-\psi\beta)^{-1}\ \underline{Z} + \sum_{j=2}^{m} \psi\beta\lambda_j(1-\psi\beta\lambda_j)^{-1}\underline{V}_j. \quad (11.32)$$

Using equation (11.32) in equation (11.17) enables the general model to be written as

$$\underline{p} = [\psi(1-\beta)\underline{b}\ \underline{A} + (1-\psi)\underline{h}][\underline{I} + \psi\beta(1-\psi\beta)^{-1}\underline{Z} + \sum_{j=2}^{m} \psi\beta\lambda_j(1-\psi\beta\lambda_j)^{-1}\underline{V}_j], \quad (11.33)$$

which using equations (11.21) and (11.22) simplifies to

$$\underline{p} = \psi\beta\tilde{\underline{p}} + [\psi(1-\beta)\underline{b}\ \underline{A} + (1-\psi)\underline{h}][\underline{I} + \sum_{j=2}^{m} \psi\beta\lambda_j(1-\psi\beta\lambda_j)^{-1}\underline{V}_j] \quad ,$$

$$= \hat{\underline{p}} + [\psi(1-\beta)\underline{b}\ \underline{A} + (1-\psi)\underline{h}][\sum_{j=2}^{m} \psi\beta\lambda_j(1-\psi\beta\lambda_j)^{-1}\underline{V}_j]. \quad (11.34)$$

The second term on the second line of equation (11.34) clearly gives the difference $\underline{p} - \hat{\underline{p}}$ and this is the percent deviation of $\underline{p}$ from the steady state distribution $\hat{\underline{p}}$. Equations (11.33) and (11.34) represent a new decomposition of the traditional linear urban model, and the same logic can be easily transferred to any such model in which the transformation of one endogenous activity into another can be separated into a scale and distribution effect. This of course limits the usefulness of spectral decomposition for input-output analysis but it is highly relevant to urban models such as the Lowry model which have the separability property. Equations for $\underline{e}$ analogous to (11.33) and (11.34) can also be derived and it is possible to derive dual relationships between the spectral sets of $\underline{B}\ \underline{A}$ and $\underline{A}\ \underline{B}$. As these are not of central relevance here, they will not be formally presented.

The decomposition in equation (11.34) can now be simplified as follows. First set the inputs and the deviation from the steady state matrix $\underline{Z}$ as

$$\underline{w} = \psi(1-\beta)\underline{b}\ \underline{A} + (1-\psi)\underline{h} \quad , \qquad \text{and}$$

$$\underline{\Sigma} = \sum_{j=2}^{m} \psi\beta\lambda_j(1-\psi\beta\lambda_j)^{-1}\underline{V}_j \quad .$$

Equation (11.34) now becomes

$$\underline{p} = \psi\beta\tilde{\underline{p}} + \underline{w} + \underline{w}\ \underline{\Sigma} \quad , \tag{11.35}$$

which is referred to as the canonical form of the linear urban model. It is the sum of a steady state effect $\psi\beta\tilde{\underline{p}}$, an input effect $\underline{w}$ of order $(1-\psi\beta)$, and a deviation from the steady state through compounding of inputs $\underline{w}\ \underline{\Sigma}$ .

Differences between the population distributions of the three models $\underline{p}$, $\hat{\underline{p}}$ and $\bar{\underline{p}}$ , as well as the Markov model $\tilde{\underline{p}}$ can now be stated. Then $\underline{p} - \tilde{\underline{p}}$ from equation (11.35) is given in terms of the three effects as

$$\underline{p} - \tilde{\underline{p}} = (\psi\beta-1)\tilde{\underline{p}} + \underline{w} + \underline{w}\ \underline{\Sigma} \quad ,$$

which clearly sums to zero, as the first two terms are of order $\psi\beta-1$ and $1-\psi\beta$ which cancel, and $\underline{w}\ \underline{\Sigma}$ is a deviation. The difference $\underline{p} - \hat{\underline{p}}$ is only in terms of these deviations from the steady state

$$\underline{p} - \hat{\underline{p}} = \underline{w}\ \underline{\Sigma} \quad ,$$

while $\underline{p} - \bar{\underline{p}}$ can be given in terms of the three effects

$$\underline{p} - \bar{\underline{p}} = \psi\beta\tilde{\underline{p}} + \underline{w}\ \underline{\Sigma} - \psi\beta(1-\psi\beta)^{-1}\underline{w} \quad .$$

Other differences $\tilde{\underline{p}} - \hat{\underline{p}}$, $\tilde{\underline{p}} - \bar{\underline{p}}$ and $\hat{\underline{p}} - \bar{\underline{p}}$ are just functions of the steady state and the input data for the deviations $\underline{w}\ \underline{\Sigma}$ are only associated with the full model. Analogous relationships for $\underline{e}$, $\hat{\underline{e}}$, $\bar{\underline{e}}$ and $\tilde{\underline{e}}$ can be derived in dual form or as functions of the relationships given here.

## APPLICATIONS: A COMPARISON OF MODEL TYPES.

To demonstrate the degree of spatial invariance contained in the different model structures introduced above, these models have been applied to an eight zone representation of the Melbourne metropolitan region. In this case, the observed pattern of employment is highly concentrated in the CBD and surrounding zones while the distribution of population is much more evenly spread. Basic employment (employment in primary and manufacturing industries) is more evenly spread than total employment but is concentrated in the CBD and the west of the city. External population, measured as population in public housing, is considerably more concentrated than total population, in the CBD and in the west of the city like basic employment. In this case, n=m=8 and the patterns of observed population and employment are shown in map 1 of Figure 11.1. The distribution of basic employment and external population are not shown separately but in fact their distribution is equivalent to employment in map 10 and population in map 11, both illustrated as part of Figure 11.3 presented below. This example, although at a coarse level of spatial resolution, is a reasonably realistic one in that it is typical of the differences in activity distribution characterising many western cities, and models of these cities.

As there are 11 different distributions of population and employment to compare (from 10 model types together with the observed distributions), these have been arranged in the following order, using indices u,v = 1,2,..,11 to represent the particular distribution in question. Index 1 refers to the observed distributions while indices 2,3 and 4 refer to the full model based on actual interaction: 2 refers to the model with both inputs, 3 to that with only basic employment (the Lowry model) and 4 to that with

only external population. Model 5 is that based on no exogenous inputs, that is the Markov or Coleman model. Indices 6,7 and 8 refer to the steady state model in equations (11.22) and (11.23); 6 is the full model with both inputs, 7 the model with only basic employment, and 8 the model with only external population. Finally, indices 9, 10 and 11 refer to the models based on 'no interaction', given in equations (11.24) and (11.25) Model 9 is the full model with both inputs, 10 with only basic employment and 11 with only external population. The maps 1-11 which are produced in Figures 11.1 to 11.3 below refer to population and employment distributions from each of these model types. In the sequel, any value of population $p_j$ or employment $e_i$ will be superscripted by its model type index, u,v where necessary.

As a first step in evaluating and comparing the 10 models and the observed distributions, the ratios of endogenous to exogenous activity - population and employment - associated with each model are presented in Table 11.1. This table shows immediately the differences in model structure in terms of the absence or presence of inputs and outputs as well as the overall weight of exogenous and endogenous variables in determing the ultimate distribution of population and employment. From Table 11.1, it is clear that the 10 model types cover a wide range of assumptions concerning the effect of input and output variables, from models based entirely on input data - models 9, 10 and 11 to that based on no input data but only on the effect of the spatial transformations - model 5. Note also that model 10 predicts employment as entirely basic employment, and model 11 population as entirely external population.

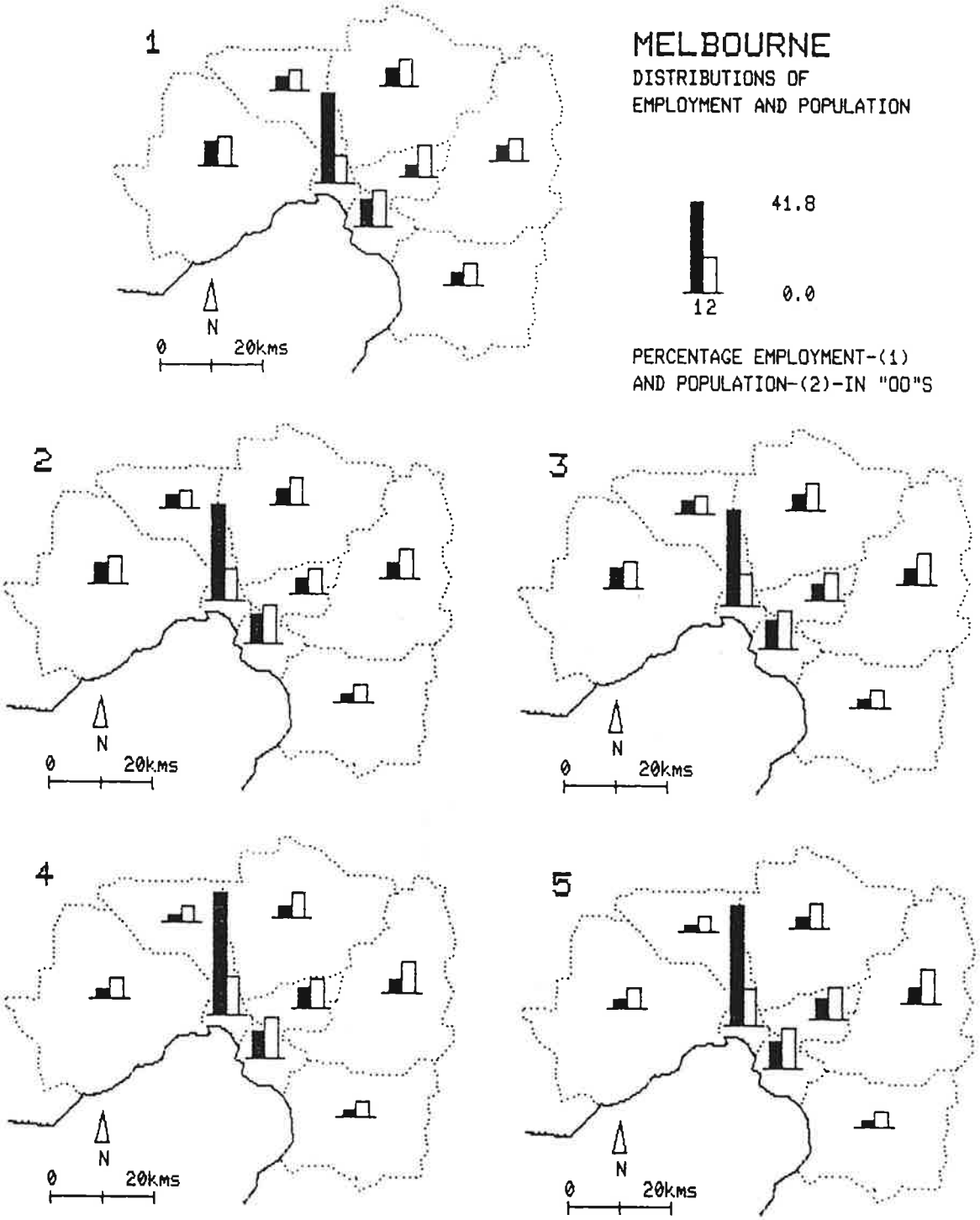
It is also possible to speculate on similarities and differences between

Table 11.1.: Classification of Model Types by Weight of Variables.

| | | | POPULATION | | | EMPLOYMENT | | |
|---|---|---|---|---|---|---|---|---|
| | | | Endogenous | Exogenous | | Endogenous | Exogenous | |
| | | | Service Population | Basic Population | External Population | Service Employment | Basic Employment | External Employment |
| MODEL TYPES, u | | | $\psi\beta$ | $\psi(1-\beta)$ | $(1-\psi)$ | $\psi\beta$ | $(1-\beta)$ | $\beta(1-\psi)$ |
| 2 | ACTUAL INTERACTION | $0<\psi\beta<1$ | 0.6153 | 0.3506 | 0.0341 | 0.6153 | 0.3629 | 0.0217 |
| 3 | | $\psi = 1$ | 0.6371 | 0.3629 | 0 | 0.6371 | 0.3629 | 0 |
| 4 | | $\beta = 1$ | 0.9659 | 0 | 0.0341 | 0.9659 | 0 | 0.0341 |
| 5 | MARKOV (Coleman) | $\psi\beta = 1$ | 1 | 0 | 0 | 1 | 0 | 0 |
| 6 | STEADY STATE INTERACTION | $0<\psi\beta<1$ | 0.6153 | 0.3506 | 0.0341 | 0.6153 | 0.3629 | 0.0217 |
| 7 | | $\psi = 1$ | 0.6371 | 0.3629 | 0 | 0.6371 | 0.3629 | 0 |
| 8 | | $\beta = 1$ | 0.9659 | 0 | 0.0341 | 0.9659 | 0 | 0.0341 |
| 9 | NO INTERACTION | $0<\psi\beta<1$ | 0 | 0.9114 | 0.0886 | 0 | 0.9436 | 0.0564 |
| 10 | | $\psi = 1$ | 0 | 1 | 0 | 0 | 1 | 0 |
| 11 | | $\beta = 1$ | 0 | 0 | 1 | 0 | 0 | 1 |

the models which will emerge in their predictions, from the prior assumptions embodied in Table 11.1. Clearly employment and population have almost identical determinants in terms of the importance of inputs and outputs, and thus it is to be expected that similarities and differences between models will be consistent in terms of population or employment. Then there are the strong similarities between the actual interaction models (2,3 and 4) and the steady state interaction models (6,7 and 8), and differences between these will be entirely in terms of the differences of the transformation matrices from their steady states. Because external population is such a small fraction of total population, models based on this as the only input (models 4 and 8) are likely to be similar in their predictions to model 5, the Coleman model, which is based on no inputs. These models too are likely to be fairly different from the others, as will be the models based on no interaction in which the inputs entirely determine the predictions (that is, models 9 and 10 which are similar in themselves and model 11).

The predicted distributions of population and employment from these models are presented in map form in Figures 11.1 to 11.3. In Figure 11.1, the observed distributions and the four models u=2, 3, 4, 5 - that is the model based on both inputs, the two models based on single inputs, and the (Markov) model based on no inputs, are presented. It is quite clear that models 2 and 3 are similar to each other and to the observed distributions, while models 4 and 5 give a much stronger concentration of employment in the CBD. However, the pattern of population generated by these models is close to the observed pattern. Figure 11.2 presents the three steady state models, models 6, 7 and 8, which on casual inspection appear close to their actual interaction equivalents, models 3, 4 and 5.




1 OBSERVED EMPLOYMENT AND POPULATION DATA

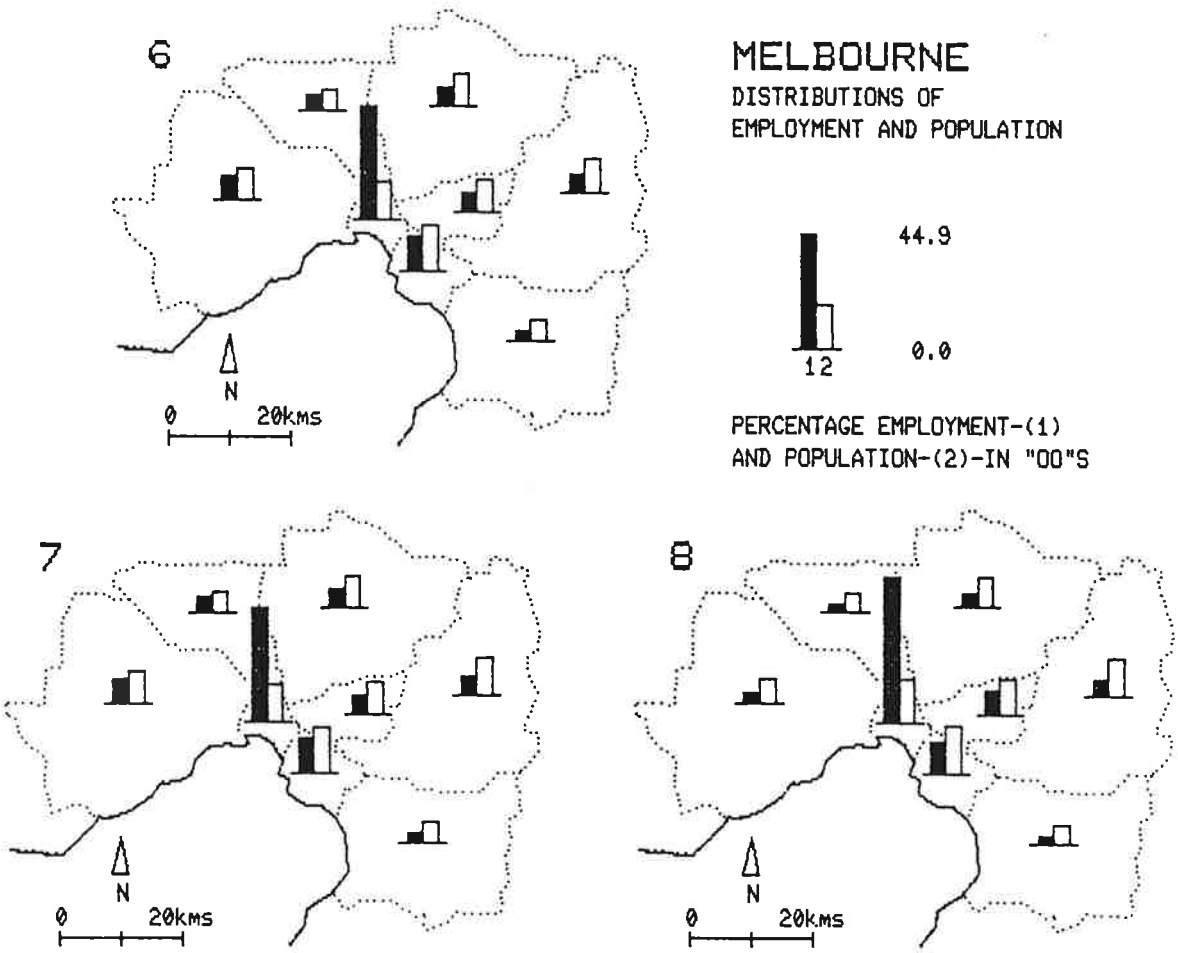
2 BASIC EMPLOYMENT AND BASIC POPULATION MODEL:OBSERVED INTERACTION

3 BASIC EMPLOYMENT MODEL:OBSERVED INTERACTION

4 BASIC POPULATION MODEL:OBSERVED INTERACTION

5 STEADY STATE MARKOV MODEL WITH NO BASIC INPUTS:OBSERVED INTERACTION

Figure 11.1: Observed and Predicted Distributions of Employment and Population for the Models Based on Observed Interaction Patterns.




6 BASIC EMPLOYMENT AND BASIC POPULATION MODEL:STEADY STATE INTERACTION

7 BASIC EMPLOYMENT MODEL:STEADY STATE INTERACTION

8 BASIC POPULATION MODEL:STEADY STATE INTERACTION

Figure 11.2: Predicted Distributions of Employment and Population for the Models Based on Steady State Interaction Patterns.

This in fact is the first indication that the effect of the transformation matrices is close to their steady state forms. Figure 11.3 presents three more extreme models, models 9, 10 and 11 based on the 'no interaction' assumption in which population and employment are direct functions of the associated input data. Models 9 and 10 generate distributions of total employment and population which are all close to the distribution of basic employment while model 11 predicts much more concentrated distributions equivalent to the distribution of external population. Although we have referred to models 9, 10 and 11 as the 'no interaction' case, this is not strictly speaking correct in that we are not assuming $\underline{A} = \underline{B} = \underline{I}$. All we assume is that $\underline{B}\,\underline{A} = \underline{I}$ and $\underline{A}\,\underline{B} = \underline{I}$, situations which can arise in many ways. However, it is possible to see the cases of actual 'no interaction' for in these cases, population in model 10 would have an identical distribution to employment and employment in model 11 an identical distribution to population.

Finally in this section, it remains to make more precise the casual comparisons emerging from Figures 11.1 to 11.3. Accordingly, we have computed precent differences between the various distributions for each pair of models. Then the percentage difference $\theta_{uv}$ between models u and v for population is given as

$$\theta_{uv} = \frac{100}{m} \sum_j \frac{|p_j^u - p_j^v|}{p_j^u}$$

and the percent difference for employment $\phi_{uv}$ is given as

$$\phi_{uv} = \frac{100}{n} \sum_i \frac{|e_i^u - e_i^v|}{e_i^u}$$

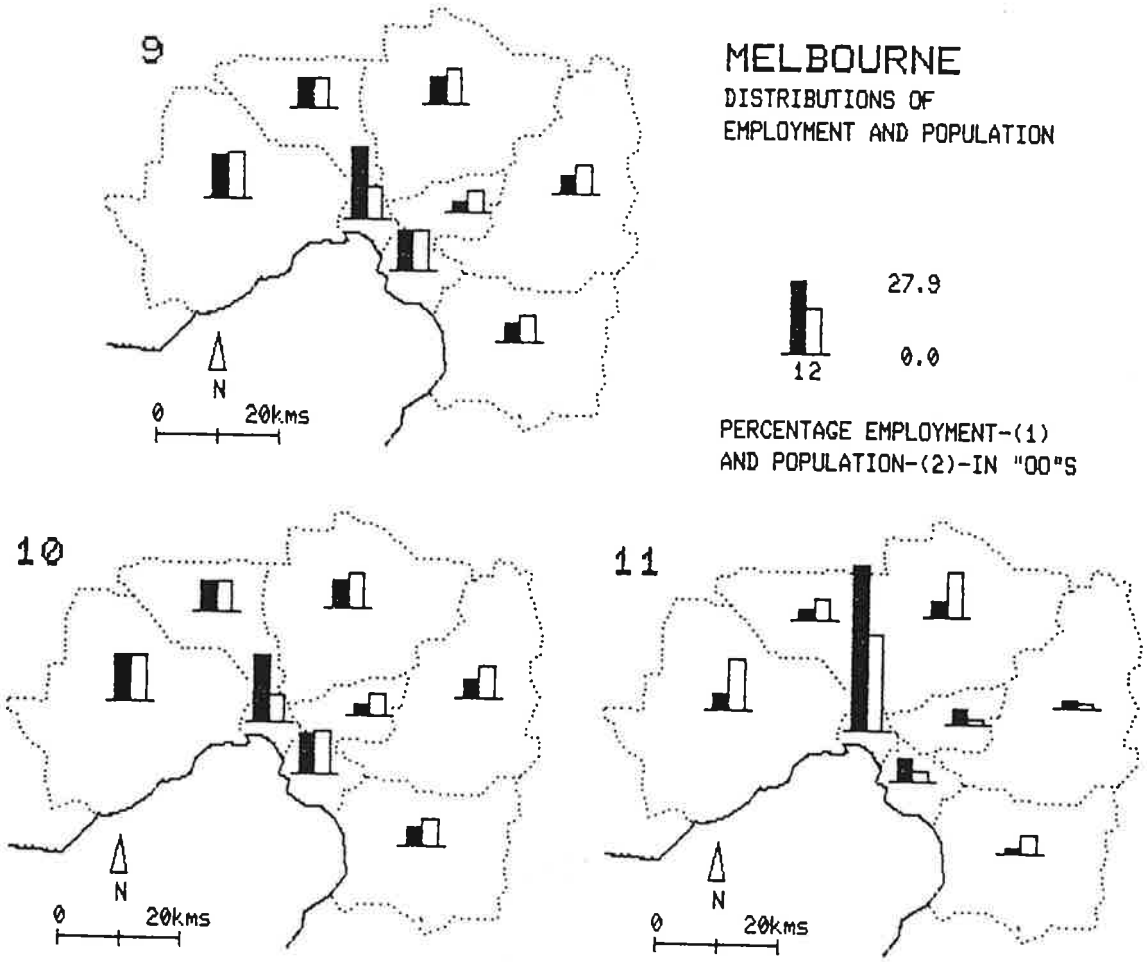



Figure 11.3: Predicted Distributions of Employment and Population for the Models Based on the 'No Interaction' Type Assumption.

These percentages are presented in Table 11.2 for the 11 distributions of population and employment respectively. These tables bear out previous observations. In terms of the observed situation, the models in which basic population is the sole determinant, and those which embody the 'no interaction' assumption perform least well. The Markov model is not close to the observed situation either but the steady state and actual interaction models where basic employment dominates, are rather close to the observed situation. This suggests that the basic employment input is a major determinant of a well-fitting model of this particular system, a point which will be made more cogent in the next section. Table 11.2 contains a large quantity of comparative information, and read with Figures 11.1 to 11.3 provides a rich source for evaluating these various models which can be further developed by the reader. Note however that the matrices in Table 11.2 are not symmetric for the base of comparison between any pair of models depends on the first model in the pair.

## SPATIAL INVARIANCE AND THE EFFECT OF MODEL STRUCTURE.

To take the analysis one stage further, it is worth exploring in quantitative terms how close the transformation matrices are to their steady states, and how the predicted distributions vary as the balance between endogenous and exogenous activity changes. Comparing the predicted distributions of population and employment for the actual interaction models with their associated steady state interaction counterparts reveals extremely small percentage differences; that is, for population and employment respectively in models 2 and 6, these are 1.48 and 2.35; for models 3 and 7, these are 1.20 and 1.37; and for models 4 and 8,

Table 11.2: Percentage Differences Between Model Types.

PERCENTAGE DEVIATIONS IN POPULATION $\theta_{uv}$

| v = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 13 | 13 | 22 | 23 | 14 | 14 | 23 | 14 | 17 | 69 |
| 2 | 13 | 0 | 3 | 14 | 15 | 1 | 4 | 21 | 21 | 21 | 62 |
| 3 | 13 | 3 | 0 | 13 | 14 | 2 | 1 | 13 | 22 | 22 | 65 |
| 4 | 24 | 15 | 14 | 0 | 3 | 13 | 13 | 1 | 38 | 38 | 68 |
| 5 | 25 | 17 | 16 | 3 | 0 | 15 | 15 | 2 | 40 | 40 | 72 |
| u = 6 | 14 | 1 | 2 | 12 | 14 | 0 | 2 | 13 | 22 | 22 | 63 |
| 7 | 14 | 4 | 1 | 12 | 13 | 2 | 0 | 12 | 23 | 23 | 66 |
| 8 | 24 | 15 | 14 | 1 | 2 | 14 | 13 | 0 | 38 | 38 | 69 |
| 9 | 17 | 21 | 22 | 34 | 36 | 22 | 23 | 35 | 0 | 6 | 64 |
| 10 | 19 | 22 | 22 | 35 | 37 | 23 | 23 | 36 | 7 | 0 | 76 |
| 11 | 156 | 156 | 162 | 178 | 186 | 159 | 165 | 179 | 127 | 139 | 0 |

PERCENT DEVIATIONS IN EMPLOYMENT $\phi_{uv}$

| v = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 13 | 13 | 39 | 39 | 15 | 15 | 39 | 30 | 34 | 42 |
| 2 | 14 | 0 | 1 | 30 | 30 | 2 | 3 | 30 | 44 | 49 | 38 |
| 3 | 14 | 1 | 0 | 30 | 30 | 2 | 2 | 30 | 44 | 49 | 38 |
| 4 | 71 | 51 | 51 | 0 | 2 | 48 | 48 | 0 | 125 | 133 | 33 |
| 5 | 72 | 51 | 51 | 2 | 0 | 49 | 49 | 1 | 127 | 134 | 35 |
| u = 6 | 16 | 2 | 2 | 29 | 29 | 0 | 1 | 29 | 46 | 51 | 37 |
| 7 | 16 | 3 | 2 | 29 | 29 | 1 | 0 | 29 | 46 | 51 | 38 |
| 8 | 71 | 51 | 51 | 0 | 1 | 48 | 48 | 0 | 126 | 133 | 34 |
| 9 | 26 | 38 | 38 | 63 | 63 | 39 | 39 | 63 | 0 | 4 | 65 |
| 10 | 30 | 42 | 42 | 67 | 67 | 43 | 43 | 67 | 4 | 0 | 69 |
| 11 | 78 | 62 | 63 | 43 | 45 | 62 | 62 | 43 | 125 | 132 | 0 |

these are 0.69 and 0.43. We can compare these with the percentage difference between the matrix $\underline{B}\ \underline{A}$ and its steady state form $\underline{Z}$ given as

$$\rho = \frac{100}{m^2} \sum_{k\ell} \frac{\left| \sum_{j=2}^{m} \lambda_j V_{k\ell j} \right|}{Z_{k\ell}}$$

In this example, $\rho$ is 31.15 which is considerably different from the ultimate percentage differences between the locational distributions.

However, it is necessary to take account of the convergence of $\underline{B}\ \underline{A}$ towards $\underline{Z}$ for only a small fraction of the difference $\underline{B}\ \underline{A} - \underline{Z}$ will be transmitted to the ultimate distributions. Thus a more useful statistic is based on the percentage differences between the actual compounded effects of $\underline{B}\ \underline{A}$ given by $\{[\underline{I} - \psi\beta\ \underline{B}\ \underline{A}]^{-1} - \underline{I}\}$ and the steady state effects given by $\psi\beta(1-\psi\beta)^{-1}\underline{Z}$. The statistic which is based on equation (11.32) is given as

$$\Omega = \frac{100}{m^2} \sum_{k\ell} \frac{\left| \sum_{j=2}^{m} \psi\beta\lambda(1-\psi\beta\lambda_j)^{-1} V_{k\ell j} \right|}{\psi\beta(1-\psi\beta)^{-1} Z_{k\ell}}$$

The value of $\Omega$ is 8.94 which implies that there is about a 9 percent difference between the actual spatial transformation of the exogenous inputs into their ultimate form and the transformation in the steady state form which is independent of such inputs. On aggregation of these differences to derive locational distributions, the percentage difference will thus be reduced to an order of 1 or 2 percent.

To complete this analysis it is worthwhile examining the same effect but excluding the actual spatial transformations contained in the eigenvectors of $\underline{B}\ \underline{A}$. Then the ratio

$$\mu = 100 \frac{\sum_{j=2}^{m} \psi\beta\lambda_j(1-\psi\beta\lambda_j)^{-1}}{\psi\beta(1-\psi\beta)^{-1}}$$

gives the value of 31.14 which is much closer to the original percentage difference between the matrices $\underline{B}\ \underline{A}$ and $\underline{Z}$. In other words, it is the similarities between the components of $\underline{V}_j$ and $\underline{Z}$ rather than their strength which determines their effect. This can also be seen by examining the vector of eigenvalues of $\underline{B}\ \underline{A}$ given as

$$[\lambda_j] = [1.00, 0.26, 0.17, 0.11, 0.09, 0.04, 0.03, 0.02]$$

where the eigenvalues are all real and positive but the ratio $\Sigma_{j=2}^{m}\lambda_j/\lambda_1$ is now of the order of 70 percent. However it is easy to see that the eigenvalues $\lambda_j$, $j \neq 1$ converge quickly towards zero in the iterative solution to the model, and measures which link these values to particular stages of the solution were used in the previous chapter to measure convergence to the invariant solution. Considerable research however remains to be done in developing this type of analysis in linear urban modelling, and this demonstration can only be regarded as a first attempt at exploring the problem.

To complete this analysis, the full model given in equations (11.13) and (11.14) has been solved for a series of values of $\psi$ and $\beta$ in the range 0 to 1. The population and employment vectors $\underline{p}$ and $\underline{e}$ from each solution have been compared with their observed distributions, with the appropriate steady state distributions $\hat{\underline{p}}$ and $\hat{\underline{e}}$ computed for each set of values of $\psi$ and $\beta$ and with the Markov solutions $\tilde{\underline{p}}$ and $\tilde{\underline{e}}$. Values of $\psi$ and $\beta$ at regular increments of 0.1 in the range 0 to 1 have been selected, thus giving 11 values of each ratio, a total of 121 varieties of each model to apply. For the actual interaction and steady state models, the extreme values of $\psi$ and $\beta$, that is $\psi = 0, 1$: $\beta = 0,1$ give the same solutions: when $\psi, \beta = 0$, $\underline{p} = \hat{\underline{p}} = \underline{h}$, $\underline{e} = \hat{\underline{e}} = \underline{b}$; when $\psi,\beta = 1$, $\underline{p} = \hat{\underline{p}} = \tilde{\underline{p}}$, $\underline{e} = \hat{\underline{e}} = \tilde{\underline{e}}$; when $\psi = 1$ and $\beta = 0$, $\underline{p} = \hat{\underline{p}} = \underline{b}\ \underline{A}$, $\underline{e} = \hat{\underline{e}} = \underline{b}$;

and when $\psi = 0$, $\beta = 1$ $\underline{p} = \hat{\underline{p}} = \underline{h}$, $\underline{e} = \hat{\underline{e}} = \underline{h}\ \underline{B}$. Thus the range over which these models are solved includes the no interaction and Markov models developed in an earlier section.

Response surfaces plotted as contours of the percentage differences between various model solutions, observed distributions and steady state model types are illustrated in Figure 11.4. In Figures 11.4 (a) and (b), these differences are shown in terms of the observed distributions and these surfaces indicate the set of values which provide a model with the closest fit to the observed distributions. For population, the best model is that with no external population and with the ratio of service to total employment about 0.4. This is close to that observed and suggests that the Lowry model is most suitable for this example. In terms of employment however, the best model is that with $\psi \simeq 0.7$ and $\beta \simeq 0.5$. We have not taken this type of analysis any further but clearly the notion of selecting the best type of model in terms of the balance of endogenous to exogenous variables is a further spinoff from developing a general framework such as this for linear urban models.

The substantial range of percentage differences between predicted and observed distributions (from 70 to 10 percent for population) is not repeated when the predictions based on actual and steady state interactions are compared in Figures 11.4 (c) and (d). Here the greatest percent difference is only about 4 percent and this bears out the fact that the input assumptions are obviously more critical in the spatial variation in the model's solutions, than the spatial transformations. Finally a comparison between model predictions and the steady state distributions from the Markov model is presented in Figures 11.4 (e)

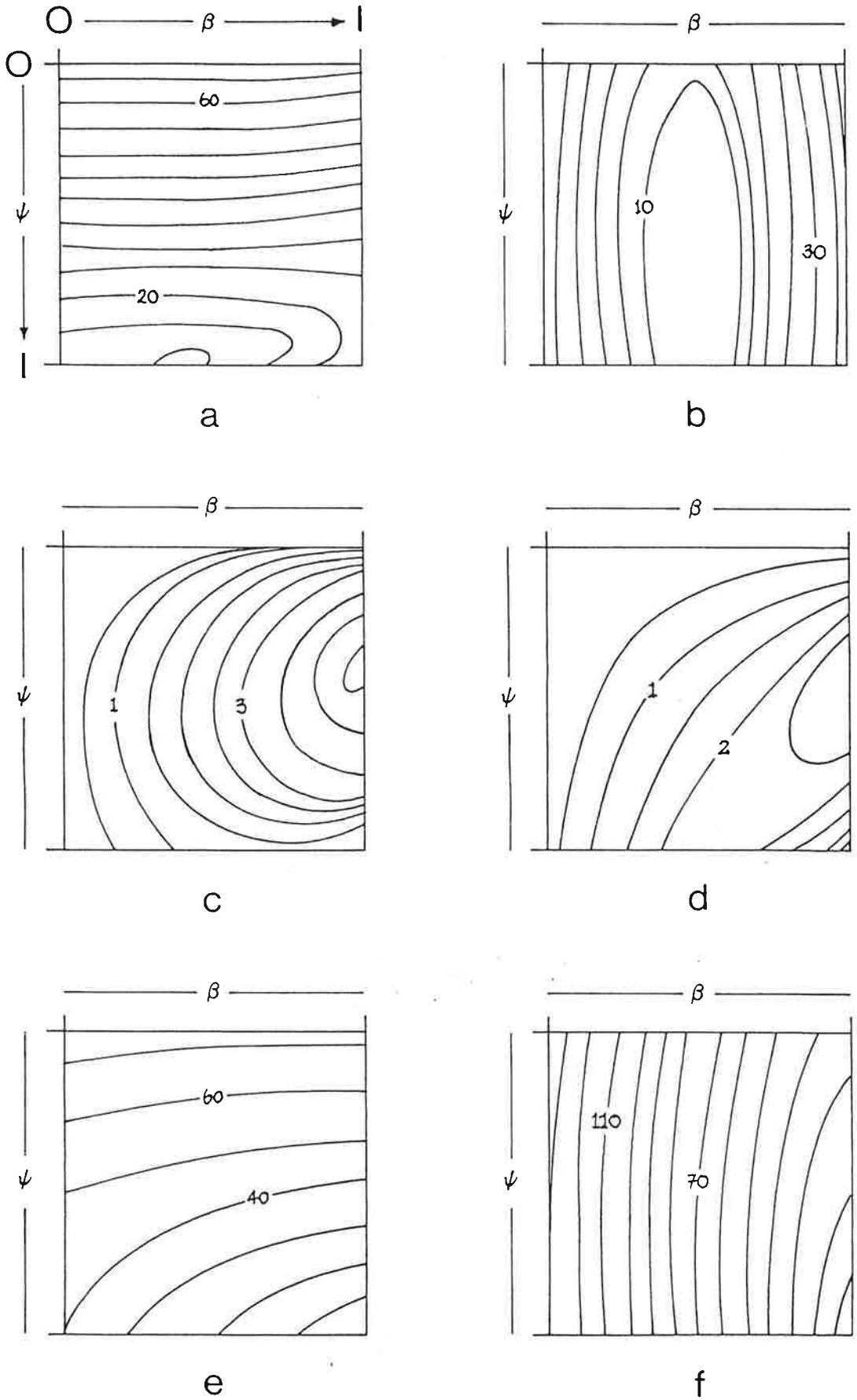



Figure 11.4: Comparisons of Model Types over the Range of Assumptions Concerning Weight of Inputs.

and (f). Here the range of variation is quite substantial (from 0 to over 130 percent) and this once again indicates that the existence of inputs is a major determinant of spatial distribution. These points have some significance for model design and these will now be developed by way of conclusion.

## CONCLUSIONS.

One obvious rule in application of the urban models developed here to spatial distributions relates to partitioning such distributions so that endogenous and exogenous population and employment distributions are radically different. This, it has been argued, will ensure a meaningful spatial transformation of activities into one another. However as shown here, if such transformations are close to their steady states, then it is the transformations rather than any input distribution which are significant. Furthermore if the input activity is only a small fraction of the total and if the transformation is near the steady state, the model is close to the endogenous Markov version. In contrast if the input data is a large fraction and the transformations far from their steady states, the solutions will be quite sensitive to the influence of both these distributions. This suggests that there is no *a priori* set of rules which indicates how much or how different input distributions should be from one another, but that the overall weight, their distribution and the nature of their spatial transformation within the model should all be considered together to judge the quality and non-triviality of such spatial models.

Finally, it is worth stressing that the analysis introduced here, particularly that involving the spectral decomposition of spatial

transformations, is very much an initial foray into the whole question of invariance in model solutions. This is part of a broader question concerned with the extent to which spatial distributions are 'averaged' on transformation, a question which hitherto has received little attention in spatial interaction modelling and which pertains to both linear and nonlinear model representations. In future work, such questions will be developed in greater detail with the ultimate intention of deriving statistics from spectral or variance analysis which will provide less ambiguous indicators of the importance of invariance than those used here. This, together with further applications of the models in the framework to nonspatial examples, represent the main lines for future research. Such fruitful directions represent a positive conclusion from the research developed in this thesis and will now form the basis for drawing together some of the themes elaborated here in a short concluding chapter.

# CHAPTER 12.

# CONCLUSIONS.

> "A man demonstrates his rationality, not by a commitment to fixed ideas, stereotyped procedures or immutable concepts, but by the manner in which, and the occasions on which, he changes those ideas, procedures and concepts"
>
> Stephen Toulmin, *Human Understanding*, Oxford University Press, 1972, p.51.

I introduced this thesis by arguing that the field of urban modelling required a stable context for research which would enable model-builders to pay attention to detail and to application rather than to develop yet more model types and styles. The Kuhnian vision of 'normal science' in which the majority of researchers work on questions of detail rather than on fundamentals has not characterised this field with the consequence that many modelling approaches remain semi-developed and have been abandoned at the first sign of difficulties over their application. With a field so dominated by different approaches, it is difficult to demonstrate the value of detailed research in its almost total absence and thus this thesis represents one of the first attempts to show how detailed research can lead to dramatic conclusions regarding traditional models. Indeed the central argument behind the techniques developed here is based on the idea that it is not the superficial, immediate and obvious difficulties with models that constitute the major problems of their credibility. These can be resolved by

direct research. It is the more subtle, hidden problems in model development which constitute the critical issues, and these can often only be identified by detailed, painstaking research.

This is a bold argument in the face of the high degree of uncertainty characterising the relevance of most social science research. In fields where there is little consensus over the key issues to be addressed, a variety of approaches might appear the most sensible strategy. Such intellectual pluralism would seem the order-of-the-day in urban modelling, thus reinforcing Toulmin's (1972) dictum that a commitment to fixed ideas is not an appropriate response in a rapidly changing field. In urban modelling, especially in the wider context of its applications in urban planning and its theoretical development through spatial economics, this poses a tantalising dilemma. To do detailed research requires a stable context, but to produce research with long-lasting relevance requires continual adaptation to change. To explore the finer points of model development which may be crucial to successful application requires a tolerance and stability of context which is all too rare in the turbulent world of planning and policy analysis where the issues attended to and the questions being asked are continually changing.

Nowhere in this thesis have I addressed this broader context in which urban modelling research takes place although I have speculated elsewhere in many papers on the issues involved (see Batty, 1978 ). To conclude this research then, the place of the ideas developed throughout the thesis will be identified in this broader context.

As noted in Chapter 1, urban models first developed in the United States in the late 1950's as a response to the demands of planners and policy-makers for appropriate techniques to analyse the urban problems of the day - suburban sprawl, inner city decline, congestion and increasing mobility. In the 1960's in Britain and a little later elsewhere the same responses to increasing urban complexity appeared. The period was short-lived especially outside the United States. In the decade of the 1970's, growth gave way to decline in economic terms and the key planning issues became those of more immediate concern - economic initiatives, unemployment, housing decay and so on. From the 1950's, planning shifted from a comprehensive, strategic focus to a narrower, more partial, pragmatic stance.

Urban models draw essentially on spatial economic theory which is mainly conceived in formal terms. Since the 1950's, formal approaches in the social sciences have been increasingly questioned as researchers in many fields have come to the conclusion that formal models are too simplistic, that social theory must encompass much richer notions of complexity which are at present, perhaps forever beyond any possibility of articulation through formal theory. In the light of this instability in both the practical and intellectual base of planning, it might seem foolhardy to argue for detailed, laborious research in a subject area which may, as many argue, be on the wrong track. However it is the argument of this thesis that only by pursuing such research to its most detailed conclusions can the abandonment of a research strategy be considered in favour of another. Moreover a related argument suggests that such research should be undertaken regardless of the general, superficial difficulties encountered because appropriate

social theory can only emerge from intellectual pluralism.

## THE NORMAL SCIENCE OF URBAN MODELLING.

Kuhn's (1970) model of scientific progress assumes that the grand strategy for a science is established rather quickly through a scientific revolution. This provides the paradigm - the unquestioned ground rules for elaboration and application of the science, which come to dominate scientific activity after the revolution. This is normal science. In time, normal science identifies anomalies in the grand scheme which mount up and may eventually threaten the paradigm itself, thus generating a further scientific revolution and so on. In this account as in most related philosophies of science, the activity of normal science is the progenitor of change.

In the social sciences, Kuhn's model has generated great interest and widespread speculation as to its applicability. As alluded to in Chapter 1, there appear to be two responses. First, that the social sciences are continually in crisis, in revolution, that there are many competing paradigms and that this marks a prescientific situation. In time, one paradigm will dominate and normal science will emerge although in Kuhn's model, it is never clear how one from several paradigms comes to dominate. The second view is based on a more sweeping modification of Kuhn's model: that the social sciences are intrinsically multi-paradigmatic and that intellectual pluralism is the name of the game.

Both approaches, in fact, assume that some kind of normal scientific

activity should take place. Indeed in parts of the social sciences, in certain branches of economics and applied psychology, there is a considerable body of normal science which seems robust enough to withstand the dictates of fashion. However it is in more peripheral, more applied areas such as urban modelling that the case for normal science must still be made and thus it seems appropriate to demonstrate how this might be so in the light of the results developed here.

This thesis has sought to examine the properties of conventional model structures such as those based on the Lowry model (Lowry, 1964), and to this end, an emphasis on the process of model solution has provided the main basis for research. Although the process of solution is analogous to a process involving real-time urban dynamics, this is too broad a comparison to enable such models to become the basis for fully-dynamic forecasting. What the emphasis on the pseudo-dynamics of model solution has led to here is a much deeper appreciation of the structural properties of traditional urban models. The early chapters of the thesis demonstrated how an analysis of model solution might enable such models to be structurally elaborated, and in fact, to enable more efficient solution; and in this research, further properties of such models were revealed.

The notion that the solution process to such models might enable insights into the empirical appropriateness of various model applications to be evaluated, and in turn guidelines for applications to be set, was never considered in early research in this field. Then, important issues in modelling related to the level of disaggregation involved and theoretical problems concerning the absence of market processes,

temporal dynamics and such like in such models. However through investigating model solution, it is now clear that questions involving covariation between inputs and outputs and the separability of spatial from activity multipliers all constitute major issues of importance in model design and application. These issues affect questions of data definition, zoning and system closure and as such, critically affect the performance and ultimate applicability of such models in planning and spatial forecasting.

This investigation into pseudo-dynamics has revealed a rich line of inquiry which has led to some rather surprising results but there are many other such lines which might equally yield important insights. Questions of the way in which activity sectors are connected to one another and the sparseness or richness of such interconnections is a major area for structural analysis of model forms which might yield useful design principles. For example, the various stages in the transport model and the way sectors inter-connect in Lowry-type models has been examined by Wilson, Coelho, Macgill and Williams (1981) using optimisation theory, and more appropriate model structures have been derived. There are also important conundrums and definitional problems to be resolved in characterising the spatial-activity system which is the subject of such models. For example, the zoning problem has only been superficially tackled to date, and questions of activity definition continually pose problems. The entire question of spatial variation and its significance has hardly been broached in such modelling and this represents a critical area for further detailed research.

More radical approaches to model structure have also yielded useful

insights too although these have rarely led to new operational techniques. The approach to dynamics developed here which is based on elaborating dynamics within the model's equilibrium conditions - *intrinsic* dynamics - can be contrasted with more radical approaches to model dynamics which involve embedding such equilibrium conditions in a wider dynamics - *extrinsic* dynamics. Developments in catastrophe and bifurcation theory are in this spirit and have yielded interesting insights into questions of discontinuity, and parameter sensitivity. Moreover, Wilson's (1981) ideas are firmly grounded in extensions to conventional models yet these are useful only as analogies, as aids to imagination rather than operational models for planning. Such work however does show promise for enabling insights into current issues such as the effect of technology change on spatial structure.

Perhaps the most impressive achievement of the last two decades in urban modelling is the synthesis in which positive and normative, linear and nonlinear modelling frameworks have been linked through optimisation theory. These achievements have influenced this research in its later stages and form part of the context and generalisations developed in Chapter 2. The link from statistical optimisation - entropy-maximising to utility-maximising, developed in the context of spatial benefit-cost analysis represents an important advance which was achieved simultaneously by a variety of researchers. Wilson and his colleagues at Leeds pioneered the approach but Leonardi at IIASA[†] Brotchie at CSIRO[††] , and Harris at the University of Pennsylvania in

---

[†] International Institute of Applied Systems Analysis, Laxenburg, Austria

[††] Commonwelath Scientific and Industrial Research Organisation, Melbourne, Australia.

Philadelphia all enabled the synthesis to take place. In one sense though, this synthesis with its emphasis on seeking consistency between predictive and prescriptive analysis came a decade too late. The problems of consistency which plagued modellers and planners using models in practice could now be resolved but the models were no longer being used and the questions they addressed - those of spatial efficiency- - no longer to the fore.

These developments in theory and method have in a sense been in the tradition of normal science but much has not been, in that the majority of modelling research has attempted to approach urban simulation rather diversely. Microsimulation approaches, *ad hoc* approaches based on decision theory, changes in the focus of interest from an activity focus - housing, retailing, industry etc. to a theme focus - energy, leisure, policy etc. have not encouraged the adaptation of more detailed research to newer contexts. Careful, patient research is urgently required to enable these diverse approaches to be thoroughly tested and to slowly build up a body of experience and expertise based on thoroughly understood theory and technique.

QUALITATIVE ANALYSIS OF QUANTITATIVE MODELS.

Two related difficulties faced by urban modellers during the last decade concern the limits to formal analysis and representation. Many key characteristics of urban systems clearly relevant to planning and policy analysis cannot be represented mathematically or even if they can, such representation is arbitrary or impossible to take

further due to lack of empirical data. The second question relates to the focus of urban models. Typically, urban models represent spatial activity distributions and explain these using rudimentary processes incorporating economic behaviour. There is a clear limit to the way in which the variety of urban problems can be framed in terms of such activities, and in the rapidly changing context of urban planning where key problems are continually changing, such modelling styles of analysis will almost inevitably fall in-and-out of fashion.

Two main responses to the 'falling-out-of fashion' syndrome have emerged in urban modelling. First there is now a strong tendency to argue that such models represent *analogues*, ways of thinking which would enable light to be cast on problems in which it is less easy to see how formal and systematic analysis might take place. Second, there is the response that models represent a useful *discipline* or set of anchor points to root the planning process in tangible technique and enable momentum in the analysis and solution of very difficult problems to be kept up.

There is in fact a third possibility which is beginning to be considered in urban modelling and which in fact represents one of the themes implicit here. This relates to the qualitative analysis of quantitative models. One of the main criticisms of traditional urban models is that they are too simple, that their treatment of complexity is superficial, that their emphasis on finding broad generalisations akin to the physical sciences is not the sort of characteristic ordering one is likely to find in social systems. What this thesis has revealed however

is that simple structures contain important relationships which in turn manifest themselves in much deeper, richer structures. The way components of simple structures combine is in itself never simple. Indeed the analysis pursued here although superficially quantitative has in fact been highly qualitative, with a concern for issues such as separability of components in structure, causal relationships between elements of structure and the way model structure is influenced by spatial form and vice versa.

One of the main difficulties of developing such models to a point where other critical characteristics of the urban system can be incorporated involves limits to quantification . However more direct qualitative analysis of model structures based on the examination of causal relationships and the extent to which qualitative approximations to such structures might be of operational use, could enable such extensions. For example, we still do not know whether simple model structures can be further simplified to the point where they can appear as causal diagrams but still remain useful and robust in planning. Research along these lines using qualitative techniques such as network analysis, analysis of dynamical behaviour of structures under perturbation and such like represent important possibilities.

This thesis has sought to develop work on model structure which should probably have taken place when such models were first developed. Notions concerning the generalisation of such models through their linear structure were mooted two decades ago but little was done. This was surprising given that many leads were present: input-output analysis, the linear modelling of economic relations and simple linear

dynamics such as that based on first-order probability processes were all well-known then, and have now formed the basis for much of this research. It would appear that such research has taken so long in coming not because the field has been ignorant of possibilities but because the research community is very small. For example, all those who began to address problems of generalising such models through input-output analysis have made major contributions although the number of such attempts has been exceedingly small.

It is just possible that the ideas in this research would not have emerged earlier if the field had been bigger and characterised by normal scientific activity. Some of the work here does build on ideas such as the synthesis through optimisation theory which clearly required time in coming. Moreover, the foray into pseudo-dynamics eventually led to ideas about model invariance, which were only further elaborated through Schinnar's (1978) work and through familiarity with the qualitative general equilibrium theory developed in a sociological context by Coleman (1973). However, generalisation of such model structures could have been effected 20 years ago after Harris (1966) and Garin (1966) had postulated the key relationships in these models to input-output analysis.

These conclusions have sketched the backcloth to this research and the types of dilemma which continue to characterise and determine research in the field. To finish, it is worthwhile briefly noting the specific lines of inquiry which emerge from this research and demand immediate attention, as well as those which represent side tracks. The most useful research in this thesis is that developed

in Chapters 2, 10 and 11 concerning generalisations of linear model structures, and the whole question of spatial invariance and the separability of spatial from activity sectors and multipliers. But Chapter 10 is built on insights which came only through the pains-taking examination of pseudo-dynamic processes begun in Chapter 3. And the kind of generalisation which began in Chapter 11 and its summary in Chapter 2 could only have been initiated from a concern for model solution, spatial invariance and separability.

The type of dynamics first developed in Chapter 3 is of less interest now. A fruitful means of elaborating model structures is contained there and in Chapter 4, but its applications in terms of improved solution procedures involving locational constraints and model parameters have in a sense been superceded by developments in the optimisation paradigm (Wilson, Coelho, Macgill and Williams, 1981). Possibly the integrated calibration technique developed in Chapters 8 and 9 is of operational use and interest but overall Chapters 3 to 9 represent the way in which the real insights in Chapters 2, 10 and 11 were derived.

There is now need for a careful demonstration project using these ideas. So much has been learnt about urban systems that the time now seems ripe for empirical application using these and many other ideas developed during the last decade. Far from abandoning these ideas, it is worthwhile continuing to develop a body of technique and experience suitable to new modelling styles and types which build on the structures presented here. The spatial-sectoral invariance-separability problem for example, is of clear importance in any spatial model application, and future research should be orientated to testing such ideas in

practice, thus emphasising the continuing cycle of theory and application through which more relevant and sound models can be developed.

## APPENDIX 1.

## DERIVATION OF THE ORIGINAL BAXTER-WILLIAMS MODEL.

Baxter and Williams (1975) developed an $\underline{\alpha}=\underline{I}$ model with a pseudo-dynamic form to speed up the calibration problem of $\underline{\alpha}=\underline{0}$ type models. Because the whole model must be computed through its iterative (dynamic) process, and because the calibration of the $\underline{A}$ and $\underline{B}$ matrices must be effected by solving nonlinear interdependent equations, the model must be run several times in an effort to meet the calibration criteria (Batty, 1976). Thus Baxter and Williams suggested that if the distribution matrices developed to meet locational constraints at each iteration, were used to reallocate all the activity generated so far, then on the t'th iteration (in time period [t:t-1]) all the activity associated with the input will be allocated using the matrices $\underline{A}(t)$ and $\underline{B}(t)$. Therefore, it would be possible to simply calibrate on $\underline{A}(t)$ and $\underline{B}(t)$ which are independent of each other, and thus approximate an interdependent problem of estimation with two independent ones. In this sense, the calibration problem would be substantially eased.

To demonstrate the original derivation of the model, define $\underline{p}(r)$, $\underline{e}(r)$ and $\underline{s}(r)$ as 1 x N vectors of the population, employment and service employment generated so far. Then the general identity

$$\underline{e}(r-1) = \underline{b} + \underline{s}(r-1), \tag{A1.1}$$

holds and population at time or iteration r is calculated from

$$\begin{aligned} \underline{p}(r) &= \underline{e}(r-1)\underline{A}(r), \\ &= [\underline{b} + \underline{s}(r-1)]\underline{A}(r) \end{aligned} \tag{A1.2}$$

Service employment is calculated in turn from population and the following recurrence relation can be derived

$$\begin{aligned} \underline{s}(r) &= \underline{p}(r)\underline{B}(r) = \underline{e}(r-1)\underline{A}(r)\underline{B}(r), \\ &= [\underline{b} + \underline{s}(r-1)]\underline{A}(r)\underline{B}(r). \end{aligned} \tag{A1.3}$$

In starting the iterative scheme implied by equations (A1.2) and (A1.3), the initial conditions are given as

$$\underline{p}(1) = \underline{e}(0)\underline{A}(1) = \underline{b}\underline{A}(1), \text{ and}$$

$$\underline{s}(0) = \underline{0}.$$

From equations (A1.1) and (A1.3), it is now clear that

$$\begin{aligned} \underline{e}(r) &= \underline{b} + \underline{s}(r), \\ &= \underline{b} + \underline{e}(r-1)\underline{A}(r)\underline{B}(r), \end{aligned} \tag{A1.4}$$

which is equation (4.19) of the main text. Baxter and Williams calibrate $\underline{A}(r)$ and $\underline{B}(r)$ to meet certain statistical criteria when all the activity has been generated at r=t. It is clear that $\underline{A}(t)$ and $\underline{B}(t)$ can be estimated independently but this does not imply that equation (A1.4) is stationary in any way.

An immediate consequence of the Baxter-Williams argument which they did not explore, relates to the requirement that equation (A1.4) be stationary, that is, that

$$\underline{e} = \underline{b} + \underline{e}\,\underline{A}(t)\underline{B}(t), \tag{A1.5}$$

where $\underline{e}$ is independent of r. Then once $\underline{A}(t)$ and $\underline{B}(t)$ have been estimated, equation (A1.5) can be solved directly from

$$\underline{e} = \underline{b}[\underline{I}-\underline{A}(t)\underline{B}(t)]^{-1}, \tag{A1.6}$$

which has the same structure as equations (4.12) and (4.17) in the main

text. The procedure for reaching equation (A1.6) might be as follows. First calibrate $\underline{A}(t)$ and $\underline{B}(t)$ on equation (A1.4) with r=t. Then apply equations (A1.4) with $\underline{A}(t)\underline{B}(t)$ independent of r for n further iterations, thus yielding

$$\underline{e}(r+n) = \underline{b}\{\underline{I} + \sum_{v=1}^{n} [\underline{A}(t)\underline{B}(t)]^{v} + \underline{e}(r-1)[\underline{A}(t)\underline{B}(t)]^{n+1}\}. \quad (A1.7)$$

When n=T+1, it is clear that the term involving $\underline{e}(r-1)$ is close enough to $\underline{0}$ to ignore, and thus equation (A1.7) can be approximated by equation (A1.6). The equilibrium vector $\underline{e}$ can be solved for directly from equation (A1.6) or iteratively from equation (A1.7). This argument is similar to the one which yields equation (4.22) in the main text.

## APPENDIX 2.

# TRANSIENT BEHAVIOUR OF THE MEAN TRIP LENGTH PREDICTIONS IN IN A LOWRY ( $\underline{\alpha}=\underline{0}$ ) MODEL.

The object of this appendix is to analyse the changes in the mean trip length statistics $\Delta\bar{C}(r)$ and $\Delta\bar{S}(r)$ in Lowry ($\underline{\alpha}=\underline{0}$)models, and to demonstrate formally that changes in these statistics are due to the nature of the sequential (pseudo-dynamic) process. In other words, it is assumed that there are no movers $\underline{\alpha}(r,r-u)=\underline{0}$, that the distribution matrices $\underline{A}(r) = \underline{I}\ \underline{\Lambda}$ and $\underline{B}(r) = \underline{\Gamma}\ \underline{S}$ are independent of time, and that the parameters of the model are constant, $\mu_1(r) = \mu_1$, $\mu_2(r) = \mu_2$. The analysis developed here is based on the fact that such a model reflects the combination of two dynamic processes: generation and allocation. The process of generation is geometric in its convergence but this does not affect the mean trip lengths whereas the process of allocation is Markovian, and it is this process which changes the trip lengths. Because the process can be viewed as strongly ergodic, many of the results from Markov chain theory can be used to examine its transient behaviour. But first it is necessary to state the model before its dynamic structure is explored.

Given the above assumptions, the model in equations (6.2) and (6.4) of the main text simplifies to

$$\underline{p}(r) = \Delta^*\underline{p}(r) + \sum_{w=t-T}^{r-1} \underline{p}^s(r,w), \tag{A2.1}$$

$$\underline{e}(r) = \Delta^*\underline{s}(0) + \Delta^*\underline{s}(r) + \sum_{w=t-T}^{r-1} \underline{s}^s(r,w), \tag{A2.2}$$

and it is obvious that the process terminates at time t for there are no movers. The stayers in equations (A2.1) and (A2.2) also simplify because of the absence of movers

$$\underline{p}^s(r,w) = \underline{p}^s(w,w) = \Delta^*\underline{p}(w),$$
$$\underline{s}^s(r,w) = \underline{s}^s(w,w) = \Delta^*\underline{s}(w),$$

and it is clear that the state equations (A2.1) and (A2.2) can be rewritten as

$$\underline{p}(r) = \sum_{\tau=t-T}^{r} \Delta^*\underline{p}(\tau), \quad \text{and} \tag{A2.3}$$

$$\underline{e}(r) = \Delta^*\underline{s}(0) + \sum_{\tau=t-T}^{r} \Delta^*\underline{s}(\tau). \tag{A2.4}$$

From the assumption relating to constancy of the distribution matrices, the submodels relating to $\Delta^*\underline{p}(r)$ and $\Delta^*\underline{s}(r)$ can be written as

$$\begin{aligned}\Delta^*\underline{p}(r) &= \Delta^*\underline{s}(r-1)\underline{T}\ \underline{\Lambda} \\ &= \Delta^*\underline{s}(0)[\Lambda\Gamma]^{r-t+T}[\underline{T}\ \underline{S}]^{r-t+T}\underline{T}\ \underline{\Lambda}\ ,\end{aligned} \tag{A2.5}$$

$$\begin{aligned}\Delta^*\underline{s}(r) &= \Delta^*\underline{p}(r)\underline{\Gamma}\ \underline{S} \\ &= \Delta^*\underline{s}(0)[\underline{\Lambda}\ \underline{\Gamma}]^{r-t+T+1}[\underline{T}\ \underline{S}]^{r-t+T+1}.\end{aligned} \tag{A2.6}$$

From equations (A2.5) and (A2.6), it is clear that the value of any activity in any time period can be determined as a simple function of the input which is constant and given. Thus any changes which arise in this process are due to the process itself.

As in Chapter 7, it is easier to revert to non-matrix presentation. First define any element of the matrix $\underline{V} = \underline{T}\,\underline{S}$ as

$$V_{ik} = \sum_j t_{ij} s_{jk} , \tag{A2.7}$$

and the n'th power of this matrix $\underline{V}$ as

$$V_{ij}^{(n)} = \sum_k V_{ik}^{(n-1)} V_{kj} . \tag{A2.8}$$

It is clear that $\underline{V}$ is also row stochastic as is the n'th power of $\underline{V}$ from the well-known rule that the product of two stochastic matrices is also a stochastic matrix. Next assume that the process of simulation beginning at r=t-T starts at iteration 1: thus t-T=1 and the process ends at r=T+1.

With these definitions, it is now feasible to examine the behaviour of the mean trip lengths $\Delta\bar{C}(r)$ and $\Delta\bar{S}(r)$. From equations (A2.5) and (A2.6), it is possible to explicitly compute the change in work trips $\Delta^* T_{ij}(r)$ and the change in service demands $\Delta^* S_{jk}(r)$. Then

$$\Delta^* T_{ij}(r) = (\gamma\lambda)^{r-1} \Delta^* S_i(0) \sum_\ell V_{i\ell}^{(r-1)} t_{\ell j}, \tag{A2.9}$$

and the change in population is calculated in the normal manner

$$\Delta^* P_j(r) = \lambda \sum_i \Delta^* T_{ij}(r) . \tag{A2.10}$$

The service demands are derived in analogous fashion as

$$\Delta^* S_{jk}(r) = \lambda(\gamma\lambda)^{r-1} \sum_i \Delta^* S_i(0) \sum_\ell V_{i\ell}^{(r-1)} t_{\ell j} s_{jk} , \tag{A2.11}$$

and the change in employment as

$$\Delta^* S_k(r) = \gamma \sum_j \Delta^* S_{jk}(r). \tag{A2.12}$$

Note that equations (A2.9) to (A2.12) are purely functions of the input variables which are known prior to the simulation.

Using equations (A2.9) to (A2.12), it is possible to derive equations for $\Delta\bar{C}(r)$ and $\Delta\bar{S}(r)$ purely in terms of the input variables. Then for $\Delta\bar{C}(r)$

$$\Delta\bar{C}(r) = \sum_{ij}\sum \Delta^* T_{ij}(r) c_{ij} / \sum_{ij}\sum \Delta^* T_{ij}(r) ,$$

$$= \frac{\sum_i \Delta^* S_i(0) \sum_\ell V_{i\ell}^{(r-1)} \sum_j t_{\ell j} c_{\ell j}}{\sum_i \Delta^* S_i(0)} ,$$

$$= \frac{\sum_i \Delta^* S_i(0) \sum_\ell V_{i\ell}^{(r-1)} \bar{c}_\ell}{\sum_i \Delta^* S_i(0)} . \qquad \text{(A2.13)}$$

The equation for $\Delta\bar{S}(r)$ can be stated in an analogous manner

$$\Delta\bar{S}(r) = \sum_{jk}\sum \Delta^* S_{jk}(r) c_{jk} / \sum_{jk}\sum \Delta^* S_{jk}(r),$$

$$= \frac{\sum \Delta^* S_i(0) \sum_\ell V_{i\ell}^{(r-1)} \sum_j t_{\ell j} \sum_k s_{jk} c_{jk}}{\sum_i \Delta^* S_i(0)} ,$$

$$= \frac{\sum_i \Delta^* S_i(0) \sum_\ell V_{i\ell}^{(r-1)} \sum_j t_{\ell j} \bar{s}_j}{\sum_i \Delta^* S_i(0)} . \qquad \text{(A.14)}$$

It is clear that these trip lengths depend upon the matrix $\underline{V}^{(r)}$ which is $[\underline{T}\ \underline{S}]^r$, and therefore the analysis of changes in these trip lengths can be restricted to examining the behaviour of $\underline{V}^{(r)}$ as r gets large.

The matrix $\underline{V}$ is row stochastic and as such, it can be regarded as a transition probability matrix. Moreover, $\underline{V}$ must be strongly connected in the graph theoretic sense for the pattern of interaction to make any physical sense. Indeed, it is inconceivable that $\underline{V}$ can be anything but strongly connected because urban systems which are disconnected or unilaterally connected imply triviality from an urban modelling point of view. Thus $\underline{V}$ defines the transition probability of a strongly-connected

ergodic Markov chain and successive powers of $\underline{V}$ give the probability that starting in origin i, activity will locate in destination j.

It is well-known that such a process converges to a steady state which is independent of the starting position. Formally

$$\lim_{r\to\infty} V_{ik}^{(r)} = v_k \quad , \tag{A2.15}$$

where $v_k$ is the steady-state probability and

$$\sum_k v_k = 1.$$

$v_k$ is the k'th element of the steady-state vector and each row of the steady-state matrix $[V_{ik}^{(r)}]$ in equation (A2.15) is equal to the steady-state vector. Note also that any transformation of the steady-state vector through multiplication by another stochastic matrix yields another steady-state vector. Then

$$\lim_{r\to\infty} \Sigma V_{i\ell}^{(r)} t_{\ell j} = \sum_\ell v_\ell t_{\ell j} = h_j \quad , \tag{A2.16}$$

where $h_j$ is normalised as

$$\sum_j h_j = 1.$$

Thus the vector $[h_j]$ is also unique and stochastic (Bailey, 1964).

The trip lengths in equations (A2.13) and (A2.14) can be considerably simplified in the limit using the results in equations (A2.15) and (A2.16). Then substituting equation (A2.15) into the limit of equation (A2.13) yields

$$\lim_{r\to\infty} \Delta\bar{C}(r) = \sum_\ell v_\ell \bar{c}_\ell, \tag{A2.17}$$

and equation (A2.16) into the limit of equation (A2.14) yields

$$\lim_{r\to\infty} \Delta\bar{S}(r) = \sum_{\ell} v_{\ell} \sum_{j} t_{\ell j} \bar{s}_j \ ,$$

$$= \sum_{j} h_j \bar{s}_j \ . \qquad \text{(A2.18)}$$

Equations (A2.17) and (A2.18) have an interesting interpretation: $v_\ell$ is the limit probability of employment locating in $\ell$ and this is independent of the original starting position. $h_j$ is the limit probability of population locating in j which is based on a transformation of the limit employment probability through the journey to work. These probabilities are in fact equivalent to the marginal origin probabilities of $p_{ij}(r)$ and $q_{jk}(r)$, that is

$$\lim_{r\to\infty} t_i(r) = \lim_{r\to\infty} \sum_{j} p_{ij}(r) = v_i \ , \quad \text{and}$$

$$\lim_{r\to\infty} s_j(r) = \lim_{r\to\infty} \sum_{k} q_{jk}(r) = h_j \ .$$

Equations (A2.17) and (A2.18) are thus mean trip lengths in the sense implied by equations (6.53) and (6.54) in the main text.

It is also of interest to note that this analysis is applicable to non-stationary chains where the matrix of transition probabilities is a function of an initial matrix - a prior matrix - and a constant information change matrix. In such a case where the n'th order matrix depends on the n-1'th matrix multiplied by the constant matrix, the prior effect eventually disappears and the steady-state is formed from the constant information matrix. For example, equations (5.31) to (5.34) would define such a process if the information matrices $\underline{G}(\tau)$ and $\underline{F}(\tau)$ were constant and independent of $\tau$.

Although the trip lengths converge to stable values independent of the initial configuration of input employment, the contribution each trip length makes to the cumulative trip length gets less and less. This allocation

process is embedded in a converging geometric series which controls the absolute amount of activity generated, and thus the final cumulative trip lengths depend upon the actual movement towards the steady state. Were this not the case, the calibration problem in a Lowry model could be reduced to one of finding a Markov matrix $\underline{V} = \underline{T}\ \underline{S}$ such that $\Delta\bar{C}(r) = \bar{C}$ and $\Delta\bar{S}(r) = \bar{S}$.

This problem has been studied by Bacharach (1970) following the path-breaking work on Markov decision problems by Howard (1971), and interested readers who wish to pursue the idea are referred to the review by Curry and Mackinnon (1975) for a useful survey in the urban modelling field. The cumulative trip lengths can be written as follows: on an iteration r

$$\bar{C}(r) = \sum_{\tau=0}^{r-1} [(\gamma\lambda)^{\tau}-(\gamma\lambda)^{\tau+1}][1-(\gamma\lambda)^{r}]\Delta\bar{C}(\tau), \quad \text{and} \tag{A2.19}$$

$$\bar{S}(r) = \sum_{\tau=0}^{r-1} [(\gamma\lambda)^{\tau}-(\gamma\lambda)^{\tau+1}][1-(\gamma\lambda)^{r}]\Delta\bar{S}(\tau) \ . \tag{A2.20}$$

In the limit as $r\to\infty$, the term $(\gamma\lambda)^{r}\to 0$ and it is clear that the proportion $[(\gamma\lambda)^{\tau}-(\gamma\lambda)^{\tau+1}]$ decreases as $\tau$ increases due to the fact that $(\gamma\lambda)^{\tau}>(\gamma\lambda)^{\tau+1}>(\gamma\lambda)^{\tau+2}\ldots\ldots$

Because the process is Markovian, it is one of the most regular and mathematically tractable of all stochastic processes and thus it is possible to study the convergence to the steady state using a variety of techniques based on deriving the characteristic roots or eigenvalues of the matrix $\underline{V}$. It is well-known in linear algebra theory that a square matrix of dimension N x N such as $\underline{V}$, has N eigenvalues which are not necessarily all distinct. Associated with each eigenvalue $\xi_n$, n=1,2,...,N are eigenvectors $\underline{x}_n$ and $\underline{y}_n$ compatible with the right- and left-handsides of matrix $\underline{V}$ respectively. That is,

$$\underline{V}\,\underline{x}_n' = \xi_n \underline{x}_n' \quad , \qquad \text{and} \tag{A2.21}$$

$$\underline{y}_n \underline{V} = \xi_n \underline{y}_n \quad , \tag{A2.22}$$

where the prime indicates the transpose of the appropriate row vector.

Noting that the matrices $\underline{X}$ and $\underline{Y}$ are formed from the N eigenvectors $\underline{x}_n$ and $\underline{y}_n$ respectively, it is possible to compute a diagonalised form for the transition matrix $\underline{V}$ from

$$\underline{V} = \underline{X}'\underline{\xi}\,(\underline{X}')^{-1} \quad , \tag{A2.23}$$

or from

$$\underline{V} = \underline{Y}^{-1}\underline{\xi}\,\underline{Y} \quad , \tag{A2.24}$$

where $\underline{\xi}$ is an N x N diagonal matrix of the eigenvalues $\xi_n$. If the eigenvalues are all distinct, then it is possible to manipulate equations (A2.23) and (A2.24) to give

$$\underline{V} = \underline{X}'\underline{\xi}\,\underline{Y} \ , \tag{A2.25}$$

noting that the left and right eigenvectors are orthogonal to one another, that is, $\underline{Y}\,\underline{X}' = \underline{I}$. Equation (A2.25) can be further simplified as

$$\underline{V} = \sum_{n=1}^{N} \xi_n \underline{M}_n \quad , \tag{A2.26}$$

where the matrix $\underline{M}_n$ is defined as

$$\underline{M}_n = \underline{x}_n'\,\underline{y}_n .$$

Bailey (1964) refers to equation (A2.26) as the spectral decomposition of the matrix $\underline{V}$ and its real value in the study of the convergence of the Markov process becomes apparent when $\underline{V}$ is raised to the r'th power. Then

$$\underline{V}^{(r)} = \sum_n (\xi_n)^r \underline{M}_n \quad , \tag{A2.27}$$

and it is clear that the r'th power of $\underline{V}$ depends upon the values of the various eigenvalues.

Further results depend on certain theorems known for stochastic matrices. The Frobenius-Perron theorem proves that if $\underline{V}$ is a non-negative indecomposable (strongly-connected) matrix, the largest eigenvalue is greater than zero, and this value lies between $\min_i \sum_k v_{ik}$ and $\max_i \sum_k v_{ik}$. As $\underline{V}$ is stochastic, all row sums are equal to 1, and therefore the maximum eigenvalue must be equal to 1. Assuming that the eigenvalues are ranked in order of ascending value, that is, $\xi_1 > \xi_2 > \xi_3, \ldots$, then it is clear that in the limit

$$\lim_{r \to \infty} \underline{V}(r) = \underline{M}_1$$

$$= \underline{1}'\underline{v} \quad , \qquad \text{(A2.28)}$$

where $\underline{v}$ is the steady state vector. In fact, $\underline{v}$ is the left hand eigenvector associated with $\xi_1 = 1$, and it is clear that as all other eigenvalues are less than 1, this vector will dominate the steady state. Note that the vector $\underline{1}$ is the unit row vector. Equation (A2.27) can be substituted into equations (A2.13) and (A2.14), and each trip length can then be seen as the sum of a series of components which depend upon the structure of the spectral set associated with matrix $\underline{V}$.

This type of analysis is similar to the z-transform analysis of a discrete dynamic process (see Howard, 1971) but it is only of practical use if the path of convergence of the Markov process is sought, and this would involve computation of the eigenvalues which is often a formidable and lengthy task, for large matrices. Nevertheless, the results in this Appendix were first developed in association with the adaptive algorithm given in Chapters 6 and 7, and then led to the formal analysis of spatial invariance which is presented in the concluding chapters of this thesis. The analysis here is also presented in a slightly different form in those chapters.

# REFERENCES.


Alfeld, L.E., and Graham, A.K., 1976, *Introduction to Urban Dynamics*, Wright-Allen Press, Cambridge, Massachusetts.

Allen, P.M., Sanglier, M., Boon, F., Deneuborg, J.L., and DePalma, A., 1981, *Models of Urban Settlement and Structure as Dynamic Self-Organizing Systems*, US Department of Transportation, Systems Analysis Division, Washington DC.

Alonso, W., 1964, *Location and Land Use*, Harvard University Press, Cambridge, Massachusetts.

Ayeni, B., 1979, *Concepts and Techniques in Urban Analysis*, Croom-Helm, London.

Bacharach, M., 1970, *Biproportional Matrices and Input-Output Change*, Cambridge University Press, London.

Bailey, N.J.T., 1964, *The Elements of Stochastic Processes, with Applications to the Natural Sciences*, John Wiley, New York.

Bartholomew, D.J., 1982, *Stochastic Models for Social Processes*, John Wiley, Chichester, UK.

Batey, P.W.J., and Madden, M., 1981, Demographic-Economic Forecasting within an Activity-Commodity Framework: Some Theoretical Considerations and Empirical Results, *Environment and Planning A, 13*, 1067-1083.

Batty, M., 1976, *Urban Modelling: Algorithms, Calibrations, Predictions*, Cambridge University Press, London.

Batty, M., 1978, Urban Models in the Planning Process, in D.T. Herbert and R.J. Johnston (Editors), *Geography and the Urban Environment: Volume 1: Progress in Research and Applications*, John Wiley, Chichester, UK, pp. 63-134.

Batty, M., 1981a, Symmetry and Reversibility in Social Exchange, *Journal of Mathematical Sociology, 8*, 1-41.

Batty, M., 1981b, Urban Models, in N. Wrigley and R.J. Bennett (Editors), *Quantitative Geography: A British View*, Routledge and Kegan Paul, London, pp. 181-191.

Batty, M., Bourke, R., Cormode, R., and Anderson-Nicholls, M., 1974, Experiments in Urban Modelling for County Structure Planning, *Environment and Planning A, 6*, 455-478.

Batty, M., and March, L., 1976, The Method of Residues in Urban Modelling, *Environment and Planning A, 8,* 189-214.

Batty, M., and March, L., 1978, Dynamic Urban Models based on Information-Minimising, in R.L. Martin, N.J. Thrift and R.J. Bennett (Editors), *Towards the Dynamic Analysis of Spatial Systems,* Pion, London, pp. 127-155.

Baxter, R., and Williams, I., 1975, An Automatically Calibrated Urban Model, *Environment and Planning A, 7,* 3-20.

Beaumont, J.R., and Clarke, M.C., 1980, Improving Supply Side Representations in Urban Models, with Specific Reference to Central Place Theory and Lowry Models, *Sistemi Urbani, 1,* (Aprile 1980), 3-12.

Beckman, M.J., 1974, Entropy, Gravity and Utility in Transportation Modelling, in G. Menges (Editor), *Information, Inference and Decision,* D. Reidel Publishing Company, Dordrecht, Holland, pp. 155-163.

Berechman, J., 1976, Interfacing the Urban Land-Use Activity System and the Transportation System, *Journal of Regional Science, 16,* 183-194.

Bertuglia, C.S., and Leonardi, G., 1979, Dynamic Models for Spatial Interaction, *Sistemi Urbani, 2,* (Agosto 1979), 3-25.

Bertuglia, C.S., and Leonardi, G., 1980a, A Model for the Optimal Location of Multi-Level Services, *Sistemi Urbani, 2-3,* (Dicembre 1980), 283-297.

Bertuglia, C.S., and Leonardi, G., 1980b, Heuristic Algorithms for the Normative Location of Retail Activities Systems, *Papers of the Regional Science Association, 44,* 149-159.

Bertuglia, C.S., Occelli, S., Rabino G., and Tadei, R., 1980, A Model of Urban Structure and Development of Turin: Theoretical Aspects, *Sistemi Urbani, 1,* (Aprile 1980), 59-90.

Broadbent, T.A., 1973, Activity Analysis of Spatial-Allocation Models, *Environment and Planning, 5,* 673-691.

Brotchie, J.F., and Lesse, P.F., 1979, A Unified Approach to Urban Modelling, *Management Science, 25,* 112-113.

Brotchie, J.F., Dickey, J.W., and Sharpe, R., 1980, *TOPAZ-General Planning Technique and its Applications at the Regional, Urban and Facility Planning Levels,* Lecture Notes in Economics and Mathematical Systems 180, Springer-Verlag, Berlin.

Burdekin, R., 1979, A Dynamic Spatial Urban Model: A Generalization of Forrester's Urban Dynamics Model, *Urban Systems, 4,* 93-120.

Cesario, F.J., 1973, Parameter Estimation in Spatial Interaction Modelling, *nvironment and Planning, 5,* 503-518.

Chow, G.C., 1975, *Analysis and Control of Dynamic Economic Systems,* John Wiley, New York.

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

Coleman, J.S., 1973, *The Mathematics of Collective Action,* Heinemann Educational Books, London.

Cordey-Hayes, M., 1972, Dynamic Frameworks for Spatial Models, *Socio-Economic Planning Sciences, 5,* 73-95.

Cripps, E.L., and Foot, D.H.S., 1969, A Land-Use Model for Subregional Planning, *Regional Studies, 3,* 243-268.

Curry, L., and MacKinnon, R.D., 1975, Aggregative Dynamic Urban Models Oriented Towards Policy, C.75.12, External Research, Ministry of State, Urban Affairs, Ottawa, Canada.

Dorfman, R., Samuelson, P.A., and Solow, R.M. 1958, *Linear Programming and Economic Analysis,* McGraw-Hill Book Company, New York.

Echenique, M., 1977, An Integrated Land Use and Transport Model, *Transactions of the Martin Centre, 2,* 195-230.

Echenique, M., Crowther, D., and Lindsay, W., 1969, A Spatial Model of Urban Stock and Activity, *Regional Studies, 3,* 281-312.

Echenique, M., Feo, A., Herrera, R., and Riquezes, J., 1974, A Disaggregated Model of Urban Spatial Structure: Theoretical Considerations, *Environment and Planning A, 6,* 33-63.

Erlander, S., 1980, *Optimal Spatial Interaction and the Gravity Model,* Lecture Notes in Economics and Mathematical Systems 173, Springer-Verlag, Berlin.

Evans, A.W., 1970, Some Properties of Trip Distribution Models, *Transportation Research, 4,* 19-36.

Evans, A.W., 1971, The Calibration of Trip Distribution Models with Exponential or Similar Cost Functions, *Transportation Research, 5,* 15-38.

Evans, S., 1973, A Relationship between the Gravity Model for Trip Distribution and the Transportation Problem in Linear Programming, *Transportation Research, 7,* 39-61.

Feo, A., Herrera, R., Riquezes, J., and Echenique, M., 1975, A Disaggregated Model for Caracas, in R. Baxter, M. Echenique and J. Owers (Editors), *Urban Development Models,* The Construction Press, Lancaster, UK, pp. 175-202.

Foot, D., 1981, *Operational Urban Models: An Introduction,* Methuen, London.

Forrester, J.W., 1969, *Urban Dynamics*, The MIT Press, Cambridge, Massachusetts.

Fox, L., 1964, *An Introduction to Numerical Linear Algebra*, Clarendon Press, Oxford, UK.

Gale, D., 1960, *The Theory of Linear Economic Models*, McGraw-Hill Book Company, New York.

Garin, R.A., 1966, A Matrix Formulation of the Lowry Model for Intra-Metropolitan Activity Location, *Journal of the American Institute of Planners, 32*, 361-364.

Geraldes, P., Echenique, M.H. and Williams, I.N., 1978, A Spatial Economic Model for Bilbao, A paper presented at the PTRC Annual Meeting, 1978, University of Warwick, Warwick, UK.

Goldner, W., 1974, Projective Land Use Model (PLUM), US Department of Transportation, Washington DC.

Gordon, P., and Ledent, J., 1980, Modelling the Dynamics of Metropolitan Areas: A Demoeconomic Approach, *Environment and Planning A, 12*, 125-133.

Gordon, P., and Ledent, J., 1981, Towards an Interregional Demoeconomic Model, *Journal of Regional Science, 21*, 79-87.

Hadley, G., 1962, *Linear Programming*, Addison-Wesley Publishing Company, Reading, Massachusetts.

Harris, B., 1966, Note on Aspects of Equilibrium in Urban Growth Models, Department of City and Regional Planning, University of Pennsylvania, Philadelphia, Pennsylvania.

Harris, B., 1979, Computer Aided Urban Planning: The State of the Art, *Sistemi Urbani, 3*, (Dicembre 1979), 55-71.

Heal, G., Hughes, G., and Tarling, R., 1974, *Linear Algebra and Linear Economics*, The Macmillan Press, London.

Herbert, J.D., and Stevens, B.H., 1960, A Model for the Distribution of Residential Activity in Urban Areas, *Journal of Regional Science, 2*, 21-36.

Hill, D.M., 1965, A Growth Allocation Model for the Boston Region, *Journal of the American Institute of Planners, 31*, 111-120.

Himmelblau, D.M., 1972, *Applied Nonlinear Programming*, McGraw-Hill Book Company, New York.

Hobson, A., and Cheng, B.K., 1973, A Comparison of the Shannon and Kullback Information Measures, *Journal of Statistical Physics, 7*, 302-310.

Howard, R.A., 1971, *Dynamic Probabilistic Systems, Volume 1*, John Wiley, New York.

Hutchinson, B.G., 1976, Land Use-Transport Models in Regional Development Planning, *Socio-Economic Planning Sciences, 10*, 47-55.

Irwin, N.A., and Brand, D., 1965, Planning and Forecasting Metropolitan Development, *Traffic Quarterly, 19,* 520-540.

Jaynes, E.T., 1957, Information Theory and Statistical Mechanics, *Physical Review, 106,* 620-630, and *108,* 171-190.

Kac, M., 1969, Some Mathematical Models in Science, *Science, 166,* 695-699.

Kendrick, D., 1976, Applications of Control Theory to Macro Economics, *Annals of Economic and Social Measurement, 5,* 171-190.

Koehler, G.J., Whinston, A.B., and Wright, G.P., 1975, *Optimisation over Leontief Substitution Systems,* North Holland Publishing Company, Amsterdam.

Kowalik, J., and Osborne, M.R., 1968, *Methods for Unconstrained Optimisation Problems,* American Elsevier Publishing Company, New York.

Kuhn, T.S., 1970, *The Structure of Scientific Revolutions,* University of Chicago Press, Chicago, Illinois.

Kullback, S., 1959, *Information Theory and Statistics,* John Wiley, New York.

LCC, 1964, *London Traffic Survey: Volume I: Existing Traffic and Travel Characteristics in Greater London,* London County Council, County Hall, London.

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

Leonardi, G., 1978a, Optimum Facility Location by Accessibility Maximizing, Instituto di Scienza dei Sistemi Architettonici e Territoriali, Facolta di Architettura, Politecnico di Torino, Turin, Italy.

Leonardi, G., 1978b, Some Mathematical Programming Ideas within a Generalized Spatial Interaction and Activity Framework, Instituto di Scienza dei Sistemi Architettonici e Territoriali, Facolta di Architettura, Politecnico di Torino, Turin, Italy.

Leonardi, G., 1981, A General Accessibility and Congestion-Sensitive Multi-Activity Spatial Interaction Model, *Papers of the Regional Science Association, 47,* 1-17.

Lesse, P.F., Brotchie, J.F., Roy, J.R., and Sharpe, R., 1978, A New Philosophy for Regional Modelling, A paper presented at the Third Meeting of the Australian and New Zealand Section of the Regional Science Association, Melbourne, Victoria.

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

Lowry, I.S., 1965, A Short Course in Model Design, *Journal of the American Institute of Planners, 31,* 158-166.

Macgill, S.M., 1975, Balancing Factor Methods in Urban and Regional Analysis, Working paper, No.124, Department of Geography, University of Leeds, Leeds, UK.

Macgill, S.M., 1976, Theoretical Properties of Biproportional Matrix Adjustments, Working Paper, No.113, Department of Geography, University of Leeds, Leeds, UK.

Macgill, S.M., 1977, The Lowry Model as an Input-Output Model and its Extension to Incorporate Full Intersectoral Relations, *Regional Studies, 12,* 337-354.

Mackett, R.L., 1981, The Impact of Transport Planning Policy in the City - A Model-Based Approach applied to Leeds,unpublished PhD Thesis, University of Leeds, Leeds, UK.

March, L., 1971, Urban Systems: A Generalised Distribution Function, *London Papers in Regional Science, 2,* 157-170.

Massey, D.B., 1973, The Basic-Service Categorisation in Planning, *Regional Studies, 7,* 1-15.

Murchland, J.D., 1966, Some Remarks on the Gravity Model of Trip Distribution and an Equivalent Maximising Procedure, LSE-TNT-38, London School of Economics, London.

Oppenheim, N., 1980, *Applied Models in Urban and Regional Analysis,* Prentice-Hall, Englewood Cliffs, New Jersey.

Perez, A., 1967, Information-Theoretic Risk Estimates in Statistical Decision, *Kybernetika, 3,* 1-20.

Rees, P.H., and Wilson, A.G., 1977, *Spatial Population Analysis,* Edward Arnold, London.

Reyni, A., 1960, On Measures of Entropy and Information, *Proceedings of the Fourth Symposium on Mathematical Statistics and Probability, 1,* 547-561.

Robillard, P., and Stewart, N.F., 1974, Iterative Numerical Methods for Trip Distribution Models, *Transportation Research, 8,* 575-582.

Rogers, A., 1971, *Matrix Methods in Urban and Regional Analysis,* Holden-Day, San Francisco, California.

Romanoff, E., 1974, The Economic Base Model: A Very Special Case of Input-Output Analysis, *Journal of Regional Science, 14,* 121-129.

Said, G.M., and Hutchinson, B.G., 1980, An Urban Systems Model for the Toronto Region: 1: Model Structure, and II: Model Calibration and Evaluation, Department of Civil Engineering, University of Waterloo Waterloo, Ontario.

Samuelson, P.A., 1948, *Foundations of Economic Analysis,* Harvard University Press, Cambridge, Massachusetts.

Sayer, R.A., 1975, Dynamic Models of Urban and Regional Systems, unpublished D.Phil Thesis, University Library, Sussex University, Falmer, Brighton, Sussex, UK.

Scarf, H., 1973, *The Computation of Economic Equilibria,* Yale University Press, New Haven, Connecticut .

Scheurwater, J., 1976, The Calibration of Spatial Interaction Models by the Newton-Raphson Method, Working Paper No.5, Instituut Voor Planologie, Rijksuniversiteit Utrecht, Utrecht, Holland.

Schinner, A.D., 1978, Invariant Distributional Regularities of Nonbasic Spatial Activity Allocations: The Garin-Lowry Model Revisited, *Environment and Planning A, 10,* 327-336.

Schlager, K.J., 1965, A Land Use Plan Design Model, *Journal of the American Institute of Planners, 31,* 103-111.

Schneider, M., 1976, A Dynamic Theory of Access and Land Development, A paper presented to the Transportation Research Board Annual Conference, Washington DC.

Shannon, C.E., 1948, The Mathematical Theory of Communication, *Bell System Technical Journal, 27,* 379-423 and 623-656.

Sharpe, R., and Karlquist, A., 1980, Towards a Unifying Theory for Modelling Urban Systems, *Regional Science and Urban Economics, 10,* 241-257.

Sharpe, R., Roy, J.R., and Taylor, M.A.P., 1982, Optimizing Urban Futures, *Environment and Planning B, 9,* 209-220.

Simmonds, D., 1976, *Nonlinear Programming for Operations Research,* Prentice-Hall, Englewood Cliffs, New Jersey.

Snickars, F., and Weibull, J.W., 1977, A Minimum Information Principle: Theory and Practice, *Regional Science and Urban Economics, 7,* 137-168.

Theil, H., 1972, *Statistical Decomposition Analysis,* North Holland Publishing Company, Amsterdam.

Toulmin, S., 1972, *Human Understanding,* Oxford University Press, Oxford, UK.

Varaprasad, N., 1980, An Interactive Strategic Model of Transport Costs and Metropolitan Population Dynamics, *Environment and Planning A, 12,* 1009-1034.

Varaprasad, N., and Cordey Hayes, M., 1982, A Dynamic Urban Growth Model for Strategic Transport Planning, *Transportation Technology and Planning, 7,* 109-120.

Varga, R.S., 1962, *Matrix Iterative Analysis*, Prentice-Hall International, London.

Webber, M.J., 1979, *Information Theory and Urban Spatial Structure*, Croom-Helm, London.

Williams, H.C.W.L., and Coelho, J., 1977, Accessibility, Spatial Interaction and the Evaluation of Land-Use Transportation Plans, Unpublished Paper, Department of Geography, University of Leeds, Leeds,UK.

Williams, H.C.W.L., and Senior, M.L., 1978, Accessibility, Spatial Interaction and the Spatial Benefit Analysis of Land Use-Transportation Plans, in A. Karlquist, L.Lundquist, F. Snickars, and J.W. Weibull, (Editors), *Spatial Interaction Theory and Planning Models*, North Holland Publishing Company, Amsterdam, pp. 253-287.

Williams, H.C.W.L., and Wilson, A.G., 1977, Dynamic Models for Urban and Regional Analysis, in T. Carlstein, D.N. Parkes and N.J. Thrift (Editors), *Timing Space and Space Time*, Edward Arnold, London.

Williams, I.N., 1979, An Approach to Solving Spatial-Allocation Models with Constraints, *Environment and Planning A, 11*, 3-22.

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

Wilson, A.G., 1970, *Entropy in Urban and Regional Modelling*, Pion, London.

Wilson, A.G., 1974, *Urban and Regional Models in Geography and Planning*, John Wiley, London.

Wlson, A.G., 1981, *Catastrophe Theory and Bifurcations: Applications to Urban and Regional Systems*, Croom Helm, London.

Wilson, A.G., 1982, The Evolution of Urban Spatial Structure: A Review of Progress and Research Problems using SIA Models, *Bulletin of the Institute of Mathematics and its Applications, 18*, 90-100.

Wilson, A.G., Coelho, J.D., Macgill, S.M., and Williams, H.C.W.L., 1981, *Optimisation in Locational and Transport Analysis*, John Wiley, Chichester, UK.

Wilson, A.G., Rees, P.H., and Leigh, C.M. (Editors), 1977, *Models of Cities and Regions*, John Wiley and Sons, Chichester, UK.

Wilson, A.G., and Senior, M.L., 1974, Some Relationships between Entropy-Maximising Models, Mathematical Programming Models and their Duals, *Journal of Regional Science, 14*, 207-215.

Young, D.M., 1971, *Iterative Solution of Large Linear Systems*, Academic Press, New York.