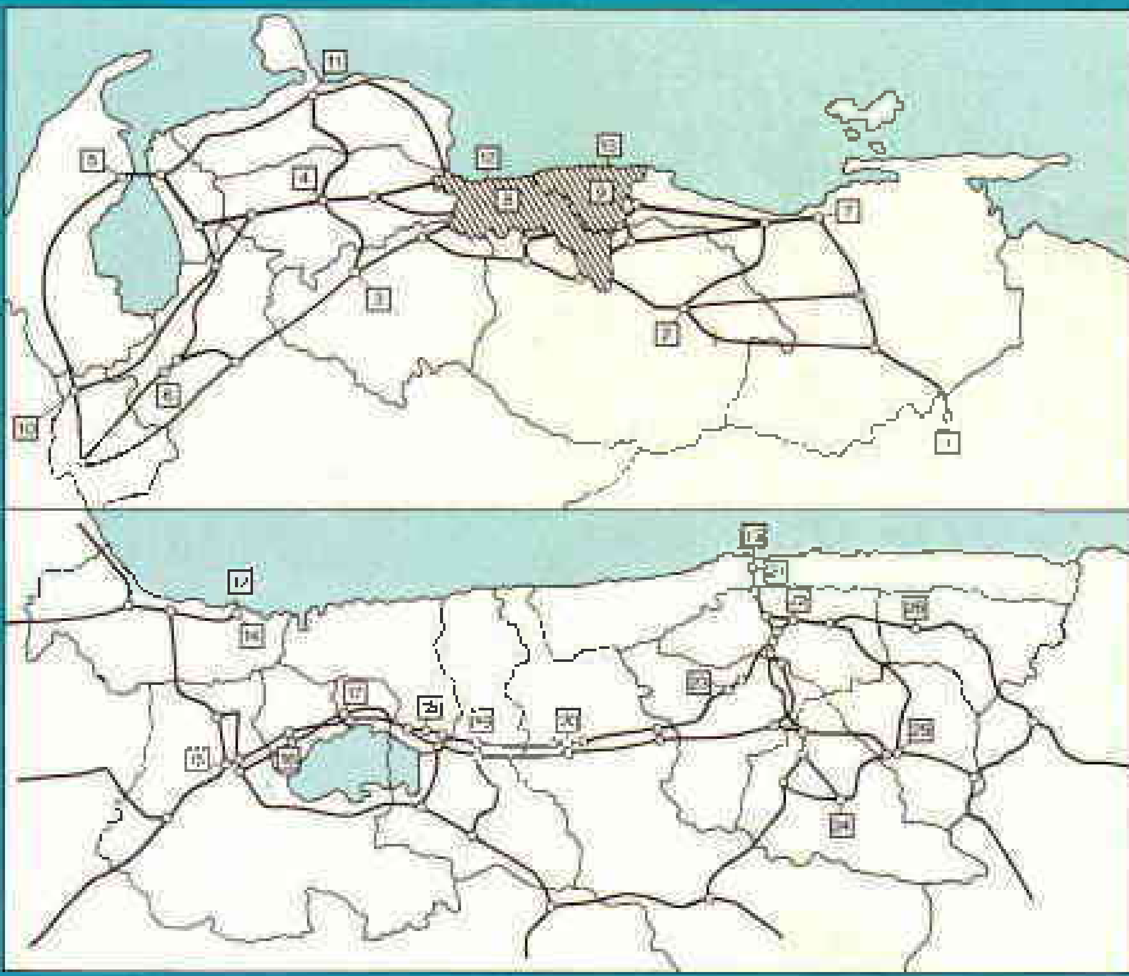
Cambridge Urban and Architectural Studies

Integrated land use and transport modelling

Decision chains and hierarchies

Tomás de la Barra



INTEGRATED LAND USE AND TRANSPORT MODELLING

Decision Chains and Hierarchies

TOMÁS DE LA BARRA

Institute of Urbanism, Central University of Venezuela

CAMBRIDGE
UNIVERSITY PRESS

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011-4211 USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1989
Reprinted 1995

Printed in Great Britain by Athenæum Press Ltd, Gateshead, Tyne & Wear

British Library cataloguing in publication data

Barra. Tomás de la
Integrated land use and transport
modelling:decision chains and
heirarchies. – (Cambridge urban and
architectural studies; v. 12).
1. Urban regions. Transport. Planning
I. Title
711'.7'091732

Library of Congress cataloguing in publication data

Barra. Tomás de la. 1946–
Integrated land use and transport modelling:decision chains and
hierarchies/Tomás de la Barra.
p. cm. – (Cambridge urban and architectural studies:12)
Bibliography
Includes index.
ISBN 0-521-24318-1
1. Transportation – Planning – Mathematical models. 2. Land use –
Planning – Mathematical models. I. Title. II. Series: Cambridge
urban and architectural studies:no. 12.
HE147.7.B37 1989
380.5'068 – dc19 88-29235

ISBN 0 521 24318 1

AP

Contents

Figures

Preface

This book has a dual character. On the one hand, it is structured in the form of a textbook. Subjects are treated in a sequence such that the reader can follow from basics to more elaborate formulations; most of the elements required for its proper comprehension are given, making it self sufficient; references are deliberately kept to a minimum. On the other hand, the book is the result of extensive research, and the knowledgeable reader will find many stimulating propositions, some more novel than others, and some probably more pertinent than others.

Integrated land use and transport modelling is an area of research that reached a high peak in Britain in the early seventies, mainly with an academic interest. Real-world applications in industrialised nations, however, have been limited mainly because of the slow rate of growth of the cities of Europe. The second half of the seventies and the first half of the eighties have seen considerable advances in the development of theories and operational models in the area of transport. Most cities in Europe and the USA regularly use stand-alone transport models for their everyday planning practice, considering the location of activities and other socio-economic variables as a relatively stable set of given inputs. The situation in third world countries, however, is quite different; since cities grow so rapidly, the interaction between the location of activities and the transport system becomes a dominant issue. It is not surprising, then, that the research contained in this book is supported by applications mostly carried out in Venezuela. It is argued that this area of research will see new light, at least in Europe, because of its potential for the evaluation of energy use in cities and regions.

The idea of writing this book originated in a PhD thesis for the University of Cambridge. The aim of the thesis and of this book is to propose a general and consistent theoretical framework for land use and transport analysis. There are many advantages in presenting a unified explanation for most of the urban phenomena, and the purpose of this book is to lay this out as clearly as possible. This book, however, differs from the original thesis in many ways. Firstly, several years of research and development in the area of integrated land use and transport modelling modified substantially many of the original propositions. Next, a large number of real-world applications of the proposed methods have lead to numerous improvements. Lastly, a decade of post-graduate teaching in the Universidad Central de Venezuela has enabled the author to present matters in different logical orders, hopefully making them more comprehensible to the reader.

The first chapter of the book is devoted to the epistemology of social sciences and to models in particular. Although it is not intended to be a full treatment of this subject, it is covered to a certain length, because there are a number of frequent misunderstandings with respect to the scientific method in social sciences. In particular, the role of models has not been covered adequately by the literature. Much more work must be devoted to this subject, mainly because social scientists are going through so many methodological changes, and models are bound to play an increasingly important role.

The second chapter deals with theories and models based on micro-economic theory. They not only represent the beginning of land use and transport modelling in historical terms, but also present basic concepts that are used in the rest of the book. A third chapter reviews spatial interaction models, from their gravitational origin to entropy maximising types. At the end of this chapter a comparison between micro-economic models and the spatial interaction approach is made, not only from a theoretical point of view, but also from the point of view of their mathematical representation, a matter of great importance.

Chapter 4 discusses random utility theory, presented as the bridge between micro-economics and spatial interaction, allowing the integration between the principles of the former and the discrete, aggregated formulation of the latter. This represents the theoretical backbone for the subjects in the remaining chapters. From this starting point, chapter 5 looks at macro-economic theories, particularly the input–output model, establishing a close relation with spatial interaction. The broad theoretical framework provided by input–output analysis can be used for the representation of cities and/or regions, and if prices and elasticities are incorporated, together with random utility theory, markets can also be represented.

Chapter 6 presents a number of issues related to land use, the location of activities in space, demand and supply of floorspace and the formation of property markets. Chapter 7 covers the transportation system, its relation to land use, and the supply/demand relationships that prevail. Only those aspects of supply that affect demand analysis are treated.

Finally, chapter 8 acts as an appendix to the theoretical chapters, presenting an operative land use and transport model that has been developed by the author and Beatriz Perez. The idea in presenting such a model, code-named TRANUS, is to illustrate the practical implications of the theory. The model is first outlined, and finally a number of specific real applications are described to further illustrate its usefulness.

In the theoretical chapters, the reader will find references to specific computer programs that have been developed by the author to illustrate particular issues. In fact, some of these programs have been used in this book to produce numerical examples. Details of the programs are included in an appendix.

Acknowledgements

The number of people I would like to thank would go well beyond the natural limitations of this book. However, a minimum list must begin with my Cambridge supervisor Marcial Echenique; Lionel March suggested the idea of producing this book and encouraged me to do so; friends also acted as sparring partners, especially Ian Williams, Tony Flowerdew, Alex Catalano, Marta Vallmitjana, Peter Rickaby and Luis Carlos Palacios. Then there are the students of the Universidad Central de Venezuela, with so many intelligent and disturbing questions. It is common practice to thank the patience and endurance of the author's wife; in my case, Beatriz Perez acted as a loving critic, and since she has participated in most of the research and development, it is very difficult for me to draw a line between my work and hers. It is also customary to thank the person who typed the manuscript; nowadays, with the advent of word processors, the distance between the author's dirty manuscript and the handsome final proof has vanished, so that I can only thank my computer for reproducing on the screen faint ideas gradually growing in body without complaints. I must finally thank William Davies, my editor at Cambridge University Press, for being cooperative for such a long period of time.

1

Social sciences and planning

The view that social sciences are very different from other forms of science is common, hence the need to develop special methods. In section 1.1 it is argued that, although there are some important differences, modern developments both in social sciences and in what are called here *natural sciences* tend to reduce them, pointing towards a common methodological framework. This idea serves as an introduction to further sections where the scientific method for the social sciences is outlined, and then is shown the role that models play in it.

1.1 Natural and social sciences: the hypothesis of convergency

Traditionally, social sciences have been considered a very special form of science, different from biology, physics or natural sciences in general. Many arguments are put forward to support this distinction. Firstly, natural phenomena are said to be *permanent*, that is, they do not change through time. For instance, water boils at 100 °C at sea level and it will continue to do so in the foreseeable future. By contrast, social phenomena are considered to be ever-changing. As a result, theories that explain natural phenomena are permanent truths or *laws of nature*. It may well be that in natural sciences theories themselves change and old ones are replaced by new ones, but once adopted, they are considered as permanent. If social phenomena are ever-changing, social theories may be valid only for a short period of time; that is, they have a particular historical reference.

Because it is generally assumed that natural phenomena are permanent, theories that successfully explain their past behaviour have strong predictive power. In social sciences, even if a theory has been very successful in explaining a phenomenon that occurred in the past, it can only provide predictions as long as the historical conditions prevail.

Another difference that is commonly pointed out is that natural phenomena can be reproduced in the controlled conditions of a laboratory, which isolates them from unwanted externalities. The use of experiments has become a key element in the creation of knowledge in natural sciences. In social sciences it is impossible to experiment in this way, because it would be expensive, unfair to the people involved, and because the mere fact that an experiment is underway affects behaviour and hence distorts results. Social sciences are restricted because empirical evidence can only be inferred from data describing past events, and therefore a particular phenomenon can never be isolated from its complex surroundings.

Furthermore, a popular argument to explain the differences between natural and social sciences is that, since the social scientist is himself part of the social environment, his propositions are influenced by ideology and by his personal relationship with the phenomena being investigated. The natural scientist, on the contrary, is said to be neutral in that his political views bear no relationship with his research.

In spite of the strength of the arguments which attempt to make a distinction between natural and social sciences, scientists in both fields are beginning to recognise that such differences are fading, making it difficult to draw a clear line of demarcation. There are many factors that have helped to bring this about. In the last few decades, natural sciences have been shaken by the introduction of theories that have completely changed, if not contradicted, old theoretical bastions. The theory of relativity is often quoted as an example of a fundamental change in scientific thinking. Newton's law of gravity was thought to be a permanent truth, verified by innumerable experiments and able to produce successful and useful predictions. Yet relativity emerged as a new theory, and to make matters worse, it was based on little empirical evidence and difficult to prove through experiments. This was a direct challenge to *induction* as the method for creating knowledge. Induction states that scientific theories have to be inferred from direct observations of reality; for relativity this was clearly not the case. It became gradually evident that laws of nature should be understood as intellectual constructions rather than substantive real entities. Scientists became sceptical about permanent truths in general.

Nowadays, theories in natural science are simply considered as a set of propositions that attempt an explanation of real phenomena. There can be a large number of theories explaining particular phenomena, so that a scientist must choose from among competing theories the one he thinks gives a better explanation. Thus, theories are never absolute truths: they are preferred options to be replaced if no longer useful. Kuhn (1962) argues that scientists usually work with an accepted body of theories or *paradigms*. This is defined as *normal* science. However, they may come across a discovery which fundamentally contradicts or modifies an existing paradigm. A revolutionary period takes place, resulting in the replacement of the old paradigm. What has happened in this century is that the speed at which this process comes about is increasing steadily, making it difficult to distinguish between normality and revolution.

There has also been a change in the way scientists have been constructing theories about natural phenomena which brings them closer to the social sciences. The strictness between cause and effect used to be a basic principle respected by all. Between the cause 100 °C and the effect boiling water, there could be no leeway, and if the theory was to be accepted or even considered, it had to correspond exactly in all cases to the highest degree of accuracy. However, in the last few decades, this principle has been relaxed. Faced with the necessity of explaining a large number of simultaneous events, natural scientists have been creating theories in which a cause produces an effect with a certain degree of probability. Classic examples are thermodynamics and mechanical statistics. This is a relatively recent development, and it must be remembered that even Einstein abhorred the idea that in nature anything could happen by chance. Today, many natural systems in physics and biology are conceived as ever-changing entities, represented through concepts such as probabilities, likelihood, unexplained or random elements, error margins, and even catastrophes and surprises.

This has meant that both natural and social sciences are increasingly adopting similar methods. Statistics has become, perhaps, the major common ground, and an exception in the sense that it is one of the few branches of mathematics specifically developed for the social sciences and later adopted by physics and biology.

Social sciences recognised very early on the changing character of real phenomena. Marx (1847) criticised the formulation of permanent social laws which classical economists had adopted so enthusiastically. Such laws, Marx claimed, must be understood as mere human constructions, therefore, they can be adopted as long as they are useful. 'The same men who establish social relations . . . also produce principles, laws and categories, in conformity with their social relations. Thus, these categories are no more eternal than the relations which they express. They are historical and transient products.'

In the case of planning, the main subject of this book, the process of convergency has been accelerated by the fact that many disciplines that originated in the tradition of natural sciences have moved into areas like economics, town planning and architecture. Transportation engineers have expanded their area of interest from traffic to the causes that generate traffic, such as the location and socio-economic characteristics of drivers. Industrial engineers have moved from the technical aspects of productive processes, to a much broader view of the social and economic environment of production. Their contribution to social sciences is significant, the best-known being linear programming, systems analysis and cybernetics. Geography has also helped in the process of convergency, firmly incorporating urban and regional analysis into its field, and introducing many methods and theories formerly applied in physical geography.

Many authors describe this process as a *quantitative revolution* (Batty, 1976), because it represents an important change in the way of thinking. Just over a century ago, the social sciences were considered by the scientific establishment as a soft area of research, interested in subjects that nobody could seriously describe as scientific problems. In the battle for recognition, social scientists reacted by building a defensive wall, insisting that exclusive methods and a jargon of its own were required. Those initiated were expected to adopt these peculiarities wholeheartedly, and any attempts to use methods from, or analogies to, natural sciences were denounced as anathema. Nowadays, as social sciences have acquired a solid reputation, this tendency to break away has lost strength, and cross-fertilisation with other disciplines is no longer considered negatively.

In this process of convergency, the invention and increasing availability of computers has become another important common element. Social events are best described as large sets of data that are easily manipulated by computers. Once data have been stored conveniently, many different types of analysis become possible, from simple statistics to complex numerical analysis and econometrics. In the case of urban and regional studies, the possibility of building large data-banks with spatial referencing enables a kind of analysis that would be impossible without a computer.

The main contribution of digital computing is, however, the possibility of *simulation*, enabling the social scientist to perform experiments. This, it was mentioned, was something that social scientists could not do in the past. Since ideal laboratory conditions cannot be created, simulation becomes a viable substitution: instead of making experiments in real environments, the analyst recreates a simplified version of reality in a computer model. Once satisfied that the model reproduces reality to an accepted degree of accuracy, the analyst can perform experiments on the artificial environment. The importance of these experiments contributes not only to decision making in planning, but to theory building in scientific research as well.

1.2 The scientific method

Here, the process of constructing scientific knowledge is presented, based primarily on Popper's (1963, 1972) explanation. These ideas are then adapted, in a later section, to the way in which people build social knowledge, to the planning process, and to the creation of knowledge in social sciences. This leads to the discussion of two particular cases: the social scientist and the planner.

For a long time it was thought that a scientist distinguished himself from other intellectuals because he used facts to support his statements about the real world. This process of building hypotheses on accumulated observations of specific instances is known as *induction*, and was originally described by Bacon. According to this method, the scientist identifies a phenomenon located at some point on the fringe between what is known and what is ignored. Scientists collect information about the phenomenon until several hypotheses can be formulated in order to explain it. Discussion then takes place and the hypothesis that manages to collect the largest number of corroborating facts becomes the accepted truth. The frontier between knowledge and ignorance is then pushed one step further, and a completely new area of research becomes available.

Induction is based on the principle that confirming evidence is the proof a particular hypothesis requires in order to be accepted, and therefore is the basis for the creation of scientific knowledge. However, it is clear that even if overwhelming confirming evidence leads us to believe that a certain hypothesis is correct, it does not necessarily demonstrate it. Popper (1963) uses the statement 'all swans are white' as a classical example to explain this; based on this hypothesis, a scientist can devote a considerable amount of energy looking for white swans, obtaining thousands of confirming events to show that the hypothesis is true. The magnitude of the confirming evidence may be such that we may be induced psychologically to believe that we have arrived at an absolute and permanent truth, and zoologists can work happily and creatively on this basis for years – until someone discovers a single black swan. The conclusion is, then, that confirming evidence does not necessarily prove anything, yet this was thought to be the basis of all scientific knowledge. This disturbing argument was first put forward by Hume, and is known as *Hume's problem*.

To solve it, Popper proposed the concept of *logical asymmetry* between verification and falsification, which simply states that even if no amount of confirming evidence can demonstrate the truth of a statement, a single piece of refuting evidence falsifies it. This, however, must not be regarded as unfortunate, because it is precisely the refuting evidence that allows the scientist to improve upon his previous statement. In the above example, the original statement can now become 'not all swans are white', or better still, the scientist can look for an improved statement that describes swans in terms other than colour, such as the genetical structure and other elements, resulting in a more general statement which at the same time contains more information. Therefore, to create knowledge, scientists should always look for refuting rather than confirming evidence.

According to Popper, the fact that a particular statement is potentially subject to refutation, establishes the *criteria of demarcation* between science and non-science. For example, the statement 'the sun will rise tomorrow between 6 and 9 a.m.' is a scientific statement, because even if it might have a high degree of probability in particular latitudes, there is always a chance that it might fail to be true. The statement will run a greater risk if it is modified to 'between 7:30 and 7:32 a.m.', but it will be more useful. For this kind of statement it is always possible to devise an experiment that can potentially put the statement at risk. By contrast, a statement like 'Buddha is eternal' might be considered by millions as an inspiring and useful idea, but since it is impossible to refute, it cannot be considered as a scientific statement.

Popper also explains the construction of knowledge as a social process, in which three *worlds* can be distinguished: world one, the world of existing material things; world two, the world of individual minds where the process of building knowledge takes place; and world three, the world of culture, books, works of art, schools, institutions, traditions, and so on, where knowledge is accumulated. World three, then, exists independently of individuals, and for this reason is called *objective knowledge*.

The scientific process is as follows: the phenomenon being investigated exists in world one; theories are devised in world two in the form of potentially refutable propositions, which are then tested against world one events; preferred theories are added to previous knowledge, becoming objective knowledge in world three. Previously accumulated objective knowledge affects the way in which individuals perceive the real world, and provides the cultural background in the process of testing and selecting a preferred theory.

Whether or not individual solutions are accepted, they become part of objective knowledge. This is where intellectual products are stored, so that others can use them in the next cycles of scientific research. A rejected theory can be as valuable as accepted ones, because eventually it may become accepted, or because parts of it can be used to assemble an improved theory, or simply because if it failed the tests, the question as to why it failed gives rise to problems that would otherwise have remained unnoticed.

The way in which the scientific method is structured is also important. The traditional inductive method is a linear process, where the starting element is observation. The next stage in the process is generalisation, from which a hypothesis is *induced* and then tested through experimentation. If the tests provide sufficient confirming evidence, the successful statement becomes knowledge.

Popper argues that simple observation of reality cannot be the starting point in the process of research. What is chosen as a subject for observation is determined by the need to solve problems. Problem identification replaces observation as the driving element in the scientific method, after which a proposed solution (new theory) leads to the deduction of testable propositions. After tests (attempted refutations), a preference must be established between competing theories. Once a preferred solution has been decided, new problems arise, beginning the process all over again. The scientific method becomes cyclic or iterative, rather than linear. The inductive method and Popper's proposition are compared in figure 1.1.

1.3 Social phenomena and social objective knowledge

The approach presented above is particularly relevant to social phenomena, and in this section a particular adaptation of Popper's scientific method is proposed. The idea of the three worlds, to start with, can be usefully applied to social events in general. All individuals, human organisations and their relationship with the environment constitute world one. At the same time, each individual member of society constitutes world two and, as such, identifies a set of problems, conditioned by his cultural, ideological and political background. This background, the storehouse of accumulated experiences of society,

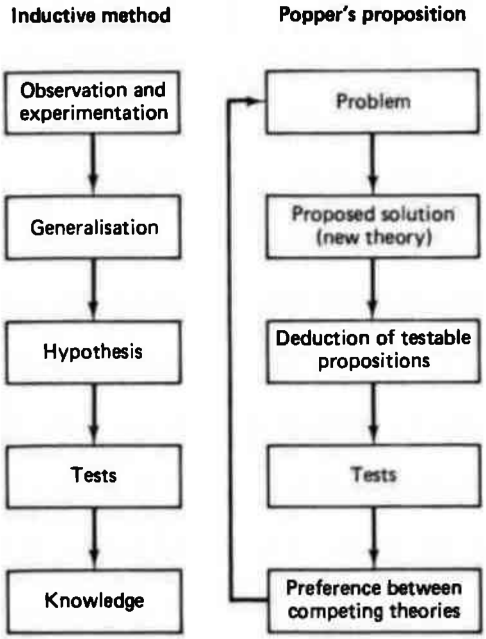


Figure 1.1. The scientific method – induction and Popper's proposition

constitutes world three, social objective knowledge. Within world two, individuals devise alternative solutions to the problems they have identified, in the form of proposed courses of action, rather than scientific theories.

If Popper's description of the scientific process is followed strictly, the next stage should be the deduction of testable propositions, that is, experiments or attempts to refute the alternative solutions. In the case of social events, however, individuals cannot perform experiments. They can, instead, resort to historical evidence by assuming that if a certain course of action did or did not work in the past, it similarly might or might not work in the foreseen future.

Each individual will decide among alternative courses of action open to him according to the level of benefit he expects from each one. In other words, individuals will estimate the benefits that each alternative course will produce, should they carry it out, and then opt for the one they think will reward them with the highest benefits. This is roughly what is called utility maximisation in economics. But how are individuals able to estimate the expected benefits? For this, they must have some idea of how society is structured, of the way in which they relate individually to the rest of society, and of how other individuals or groups will react to their own actions. Each individual will understand his social environment in a particular way, and will conceive of society as a set of elements (the state, institutions, social groups, etc.) and a set of relationships among them. In other words, each individual will have, consciously or unconsciously, a *model* of his own social environment.

Individuals will use their models of society to test the proposed courses of action in an imaginary and anticipated way. From these simulated tests, a preferred solution will emerge and action will follow, modifying world one. Action, however, may or may not produce the anticipated effects. The similitude between the anticipated effects and the outcome of the actions in real terms will depend on the accuracy of the model, on the action taken by others, and on other unpredicted events.

All courses of action will become part of world three, the objective knowledge of society. New problems will emerge, because the social environment changes, because corrections to previous actions are necessary, or as a direct consequence of the actions taken. The main feature of this process, in which a large number of individuals are simultaneously defining problems, testing alternative courses of action with numerous models, and finally acting simultaneously upon a common social environment, can be defined as the *multiplicity of social objective knowledge.*

Failure to recognise this multiplicity can lead to misconceptions. Popper (1966) himself assumes that there is a certain oneness in the creation of social knowledge, which is true, but only to a certain extent. In a way Popper makes a straightforward translation from his own description of the way in which scientists create knowledge to the way in which society builds social knowledge. In order to encourage this process, Popper (1966) proposes an *open society*, in which criticism, discussion and the principle of refutation are the means of developing social knowledge. Campbell (1969) uses a similar idea to define an *experimental society*, where real-life tests are regularly carried out to increase social knowledge. Batty (1975) describes this as a *social learning* process, but recognises the existence of conflicts. Social groups can have different, conflicting goals and thus might consider a proposed solution to a common problem as against their own interests. The role of the planner, according to this view, consists of identifying these conflicts and providing channels of communication in order to arrive at compromise solutions, halfway between two conflicting interests.

From a Marxist point of view, the possibility of achieving consensus or even compromise is limited in a class-structured society, because basic conflicts like the exploitation of labour by the owners of the means of production are unsolvable. The only possible solution is to achieve a classless society, i.e. to abolish private ownership of the means of production: only then can conflicts (these would still exist) be solved.

It can be concluded, then, that social processes are much more complicated than scientific ones because of the two factors described above: multiplicity and conflict. All members of society are actively working in numerous worlds two, so that the preferred solution in the scientific process becomes a large number of courses of action, often contradicting each other. A new social situation emerges as a result of these numerous actions, forcing their way through a network of social relationships. The degree of success of each individual will depend on his share of power within the social structure, as much as of the rightness of their actions. In order to increase their influence, individuals will tend to group with others whose model of society is similar, and with whom compromise lies within an acceptable range. This leads to political parties, trade unions, action groups, and so on. Each group will create its own objective knowledge.

1.4 Social scientists and planners

The process described above applies to all members of a particular society, but there are two special cases that deserve further discussion: the social scientist and the planner.

1.4.1 The social scientist

The aim of the social scientist is to improve knowledge about social phenomena. In order to achieve this, he will follow roughly the same process described above for all scientists.

The first question that can be raised is: to what extent is the social scientist biased in his research because of the fact that he is an active member of the society which constitutes his subject matter? The answer must be that, from the stage of problem identification to the final creation of knowledge, the social scientist is conditioned by his position in society, his cultural background and his political views. But the same can be said of any scientist. When a physicist or a biologist chooses his subject of interest, he is also conditioned by his social environment. What is certain is that the subjectivity of science in general cannot be considered as a deficiency. The fact that a scientist belongs to society does not blind him; it merely provides him with a necessary perspective, and vision and perspective go together. On the other hand, problems, whether natural or social, are problems because they affect people; if they did not, they would not be worth investigating. The only source of information available to a scientist to help him decide which are the most relevant problems and to motivate him to solve them, stems from his particular position in society.

The social scientist, then, can be seen as an active member of society investigating his own environment, in the same way as any other scientist would proceed with respect to any other subject. The fact that he belongs to society is precisely what enables him to identify problems, in this case, a particular social phenomenon of relevance. He will formulate one or more theories to explain the problem at hand, and from these he will derive testable propositions. It was said that in the case of social sciences, laboratory-type experiments cannot be performed, so that the scientist must resort to models. Models are, then, testable propositions derived from theories.

A simple example may be useful at this point. A scientist identifies education as a problem in a particular city, because he realises that students with similar IQs perform differently. He elaborates a theory that explains this as due to two factors: crowded classrooms and the educational level of parents. This theory as such cannot be tested in the traditional way, because the scientist cannot change the teacher/student ratio or the educational level of parents for the sake of the experiment. But he can build a model which states that the level of performance measured, say, in terms of success rates in examinations, is a linear function of the teacher/student ratio, and of an index of parental education. He can then collect the relevant data and test the model.

From this example, several conclusions can be drawn. Firstly, a model is a reinterpretation of a theory in a testable form. Next, it must be noted that from a single theory several models can be built. For example, it could have been an exponential function instead of a linear one, or the variables could have been measured in different terms. It is also possible that the same model can be used to test different theories, for instance, that the educational level of parents plays only a minor role. The results of the tests will help the scientist to establish a preference among the competing theories. It can also be concluded that if the tests are carried out with reference to a particular case, say the secondary schools in Bombay, the theory can be established at that level, but if a more general statement is sought, tests should be carried out with a much broader data set. Therefore, certain theories can be considered valid only for certain realities. Equally, and because of the changing character of society, theories may be valid for particular historical instances, but not for others, so that theories should be updated periodically.

1.4.2 The planner

The planner has a more direct involvement because he participates actively in the decision-making process of society. The problems he identifies might be the same as those identified by the social scientist, but the solutions he considers are courses of action rather than explanations or improved theories.

The planning process can also be viewed in a cyclical way, as shown in figure 1.2. The first step in the planning process is the identification of a problem. The planner then formulates a set of alternative courses of action in order to solve the problem. The planner will also have a theory about the way in which reality is structured, and about how it will react to the changes being considered. Planners will then test their proposed solutions, and in order to do so, they will simulate them with a model derived from the theory. The model can be expressed in a variety of forms, such as verbal, physical or mathematical. Simulation will produce as a result the probable effects of each alternative course of action.

The results of the simulations are then pre-evaluated in order to assess the positive or negative effects that each alternative course might produce on the various social groups and on the physical environment. This is called *pre*-evaluation because it takes place before action is carried out. Once alternatives have been compared, a preference must be established.

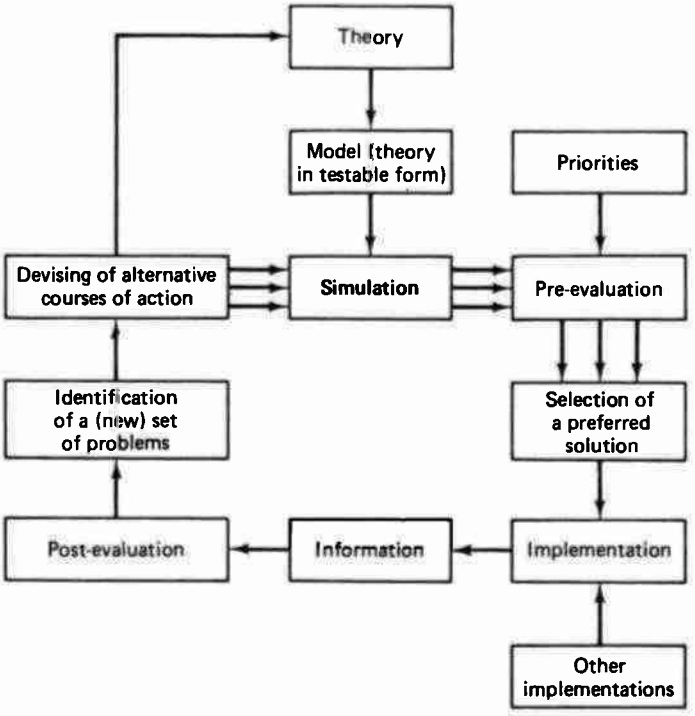


Figure 1.2. The planning system

After this, the proposed course of action is (hopefully) implemented. Once reality has been modified as a result of actions, information must be collected to monitor the way in which changes have taken place. This leads to the process of post-evaluation. New data might show that events differed from the planner's predictions, because the model and/or its underlying theory were not very realistic, because of unpredictable external events, or because other individuals or institutions implemented plans of their own. Information will also show whether identified problems have been solved; if they have not been solved completely, it will show which are the remaining problems and if new ones should be added. If there is a gap between predictions and real events, post-evaluation must determine what caused them: was it the disturbing action of others, was the model ineffective or was the theory mistaken? Usually the last two issues are taken up by social scientists.

Post-evaluation, then, can lead to the identification of a new set of problems, to improvements in the model and/or changes in the theory, thus starting a new cycle.

1.5 Some basic concepts about models

It has been argued above that models are part of all planning processes. Individuals use models in their daily decisions, and so do groups and institutions. Professional planners can be seen as individuals whose role is to help groups or institutions carry out decision-making processes. They are called upon to do this because they are supposed to have better theories about the social environment, and models which are better tests for the proposed courses of action.

Therefore, the discussion is not so much about whether or not to use models, but rather about what types of model are most suitable for particular cases. Here we begin by defining and discussing some basic concepts related to models, and particularly those most relevant to urban and regional planning. There is no intention, however, of making an exhaustive classification of models; this has been done extensively elsewhere, and useful reviews are contained in Echenique (1968), Bunge (1973), and Batty (1976).

Of necessity, all models are simplifications of reality. This is not only because it would be impossible to represent reality in all its detail: it is also intentional. The model-maker will deliberately ignore all aspects which he thinks are not essential for the problem being analysed; the model is intended to represent its *real referent* only in its most significant aspects. Hence, a model is more an idealisation of its referent, than a simplification. As Bunge (1973) puts it, 'there are as many idealisations as idealisers, data and goals. Even if two model-builders have access to the same empirical information, they may construct different models, for model building is a creative activity engaging background, abilities, and taste of the builder.'

A model has been defined as a theory in a testable form. On the other hand, a model is also an idealised representation of reality. Models are, then, an intermediary between theory and reality. General theories, being highly abstract entities unconcerned with particulars, are untestable. We have to build models which combine theoretical and real aspects. If we call T a theory, M a model, and R reality, then we can establish the relationship $T \rightarrow M \rightarrow R$.

Let us begin by discussing the $M \rightarrow R$ relationship. In general, the closer the relationship between model and reality the better, but this will depend on the purpose for which the model is built. If the model-builder's intention is to make an accurate representation of reality, then the *goodness-of-fit* between real data and simulated results becomes essential. The model-builder will be tempted to include as many elements as he can, as well as complex functional relationships, in order to represent reality in its full richness. However, if the intention is to make long-term predictions of the future behaviour of the real system, then it will be safer if only those elements which are thought to be relevant in the long run are included, overlooking short-term peculiarities.

The same principle applies if the intention of the model-builder is to prove the validity of a particular theory. If the theory is meant to stand only for specific cases, then the corresponding model should be able to represent reality in an accurate and similarly detailed fashion, resulting in a close $M \rightarrow R$ relationship. If, however, the theory is meant to be a general one to explain a large variety of real situations, then $M \rightarrow R$ cannot be close. In this case, the strength of a theory can be measured in terms of the number of different R's for which the $M \rightarrow R$ relationship approximately holds.

From the point of view of the $T \rightarrow M$ relationship, it can be said that models vary according to their degree of theoretical content. If we think of a theory as a set of statements or hypotheses that explain how a certain aspect of reality can be broken down into its components and how these components relate to each other, the theoretical content of a derived model could be measured in terms of the number of hypotheses that are made explicit. For instance, if a theory $T(H)$ contains hypotheses $H_1, \ldots, H_4$, it is possible to build a model $M(S)$, $S_1, \ldots, S_4$, where S_n denotes specific representations of the corresponding hypotheses H_n, and such a model would have a maximum degree of theoretical content. A model $M(S_1, S_3)$, say, would have a low degree of theoretical content.

As in the $M \rightarrow R$ relationship, simplicity can also favour the $T \rightarrow M$ relationship. It might be expected that an oversimplified model should fail or give only approximate results, but in fact it may tell us more about the pros and cons of the theory and of its real referent than a highly sophisticated model. In the latter case it might be difficult to obtain new insights into the problem because of the difficulty of untangling the many complex cause–effect relationships. As Bunge (1973) puts it, 'the failure of a precise idea is more instructive than the success of a muddled idea'. Oversimplified theories and models are also dangerous: they can become either hopelessly general or simply lack information.

A related concept to theoretical content is *testability*. There are some very general theories, like information theory, which are not testable unless applied to specific cases. If the theory is so general that it cannot be tested as such, it is not possible to derive from it a testable model. It can be argued that information theory is not really a theory but a method, just like statistics, geometry or systems analysis. But even if it is not possible to build testable models from such theories or methods, it is always possible to represent them as diagrams, flow charts or even complex mathematical models, which happen to be untestable models of untestable theories. To avoid confusion, such models will be called *representations*. Representations are only symbols of their theoretical referent, while models simulate real referents.

Representations can be very useful to clarify and check the consistency of a theory, but only through testable models can we actually verify its correspondence with the real referent. From this point of view, a model must be designed to contain a maximum number of outputs that can be compared against data collected from reality. In other words, the model-builder must try to maximise the possibility of failure. A risk-free model will teach us very little about a real phenomenon or the theory being tested, because it can only avoid risk by eliminating specific outputs that can be compared to reality.

Apart from the theoretical content of the model itself, the theoretical content recognised by the user has to be considered. At the lowest extreme of the scale – that is, when the user knows nothing about the internal structure of the model – the *black box* can be found. A black box is a model in which only the input and output terminals are made explicit. The user simply knows how to manipulate the inputs and how to interpret the outputs, rather like most people do when watching a television set. Inside the black box there could be models of any degree of theoretical content, but this is of no concern to the user.

From this point of view, a whole range of models can be established, from the completely black box, through grey boxes, to translucent boxes. If a black box is divided into several components linked to each other, the structure of the model becomes more explicit. What is fed into the model – the input – are generally called *exogenous variables*, that is, data that the model must accept as unquestionable facts (the selected channel on the TV). These given inputs are processed inside the box, whatever its blackness, until the results come out the other end (a picture on the screen). All variables that are calculated inside the box are called *endogenous variables*. If the box is completely black, the only endogenous variables the user will be aware of are the model outputs. If, however, the box has been divided into several parts, the user will be able to assess intermediate endogenous variables, of which only some will become final outputs.

Another key concept is that of *analogy*. If we do not know enough about some aspect of reality, it is sometimes useful to draw an analogy with another, better known, real system, about which we have a sound theory or well-tested models. In other words, in the lack of a theory that explains the behaviour of a particular phenomenon, it might be convenient to borrow a theory and/or a derived model from another system, because we think that both share important characteristics. At the beginning, the borrowed model or theory can be tested as found, from which the main differences will become apparent. This will lead to improvements, until we arrive at a completely new theory or model, which was what we were looking for in the first place. Although they have many friends and foes, analogies can be very useful in creating new knowledge, and in fact their use is probably as old as any other scientific method. It is important, however, to keep in mind that analogies are only the starting point in the development of new theories, and the sooner we depart from them the better.

Models can have a variety of forms, ranging from simple verbal statements, to diagrams, graphs, physical models (called iconic models), to mathematical models. Mathematical models, in turn, can be deterministic, probabilistic or stochastic. The deterministic model will always produce the same exact result from a given input. The probabilistic model will also produce the same results from identical given inputs, except that it will assign to them a certain degree of probability. The stochastic model will give different results each time. As will be seen later in this book, many models can have probabilities or stochastic elements as intermediate steps; probabilities can be assigned to stochastic variables, and in turn, deterministic values can be obtained from probabilities, so that the model can still produce the same results on each run, but with degrees of probabilities and randomness attached to them.

There are some rather obvious advantages in the use of highly formalised models, because they allow for more accurate predictions and more complex relationships. As was mentioned, social processes are best described as large sets of data which can only be handled through mathematics and computation. However, in urban and regional planning it is sometimes difficult to maintain a high degree of formalisation throughout an analytical process, because not enough is known about some parts of the phenomena being analysed, because of lack of data, or simply because of theoretical weakness. As a result, planners or social scientists can choose a combination of models with different degrees of formalisation. A broad mental picture of the social system as a whole can be followed by more-formalised models containing unrelated information about the system, such as maps, listings, diagrams, and so on. Some, like diagrams, can contain simple relationships, and the complexity of models can grow gradually as more relationships are established.

Model structure is also affected by the particular characteristics of the planning system. An adequate correspondence should be established between the exogenous variables of the model and the real control variables in the planning process. For example, if a planning authority has effective control only over land use, then land should be, ideally, the single exogenous variable of the model. This, however, is not always possible, but in general exogenous variables not subject to control in reality should be kept to a minimum. As will be seen later, simulation and evaluation are closely linked, and, as a consequence, what is left out of the model cannot be properly evaluated. For instance, if employment of a certain type or a particular population group is not modelled but rather given exogenously to the model, then the effects of alternative courses of action over these variables is impossible to assess, since by definition exogenous elements will not change their behaviour within the model.

1.6 Model dynamics

The choice of models is also affected by the speed of change of the real referent. If a certain element is particularly sensitive to the effect of other elements and as a consequence is highly subject to change, the corresponding model should be very dynamic with respect to that element. On the other hand, if an element of the system changes slowly, say beyond the planning horizon, then its simulation can be more static, and eventually can be given as data. As a consequence, models will not only combine different degrees of formalisation and detail, but they will also combine different dynamics, giving special emphasis to those variables that change rapidly, and treating other, more stable variables as external to the model.

A fundamental concept in system dynamics, particularly relevant to urban and regional systems, is the response time. Consider a system in dynamic equilibrium, that is, a system that will not change through time, unless affected by external stimuli at a particular time t. Figure 1.3 shows several examples of system dynamics of this kind. Figure 1.3(a) shows a case in which the change in the system occurs instantly as a result of an external factor. Figure 1.3(b) shows an example where the system takes some time to react and reaches the new steady state gradually rather than instantly. The amount of time that the system takes to reach a new equilibrium is called relaxation time, response time, or *time lag*.

External factors can take place at several times, such as at t_1, t_2 and t_3 in figure 1.3(c), and can be of different magnitudes, such as F_1, F_2 and F_3. The symbolic curve representing the behaviour of the system will

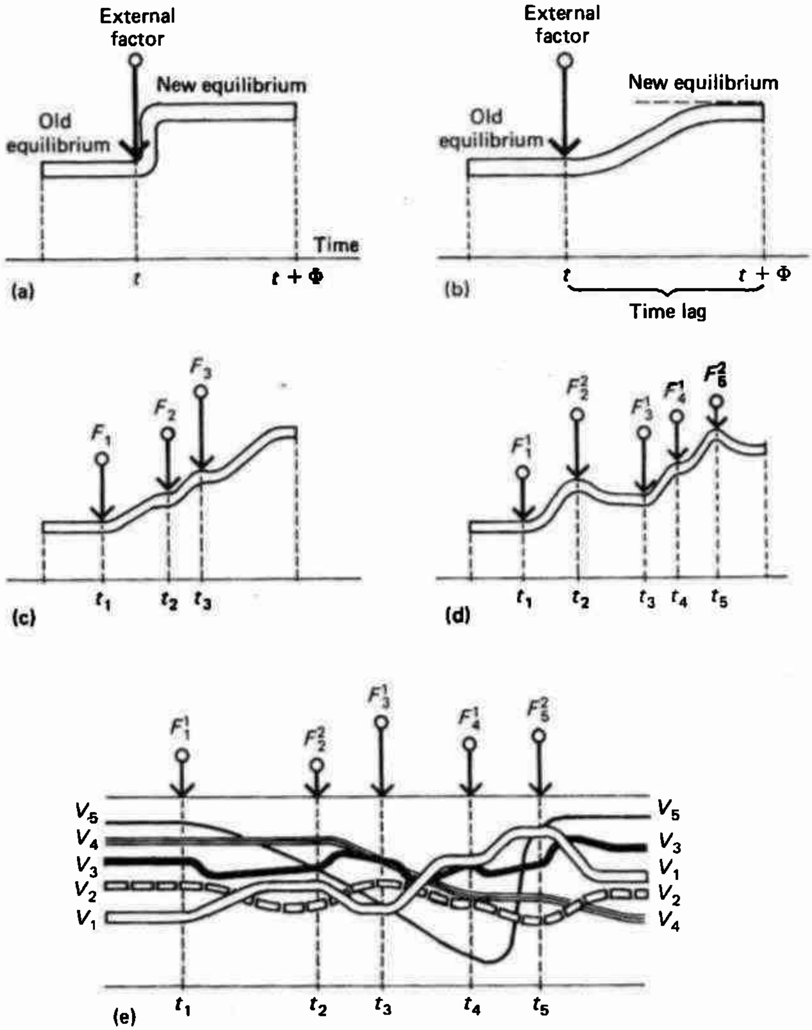


Figure 1.3. External factors affecting a system through time

enter the disrupted period from an old state of equilibrium and emerge in a new one. In this example, F_1, F_2 and F_3 are of the same nature because the effect they produce on the system is positive. The effect of different types of external factors is suggested in figure 1.3(d), where F_t^n represents a factor type n taking place at time t; in the example, factors of type $n = 1$ have a positive effect on the system, while factors of type $n = 2$ have a negative effect.

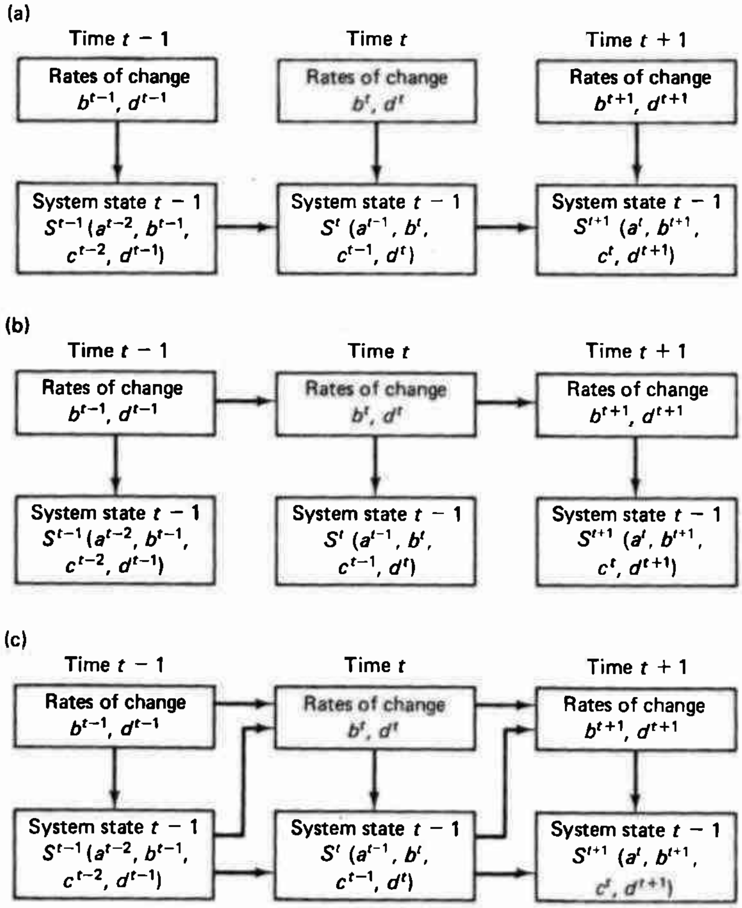
If the system is defined as a more complex set of variables, external factors of the same type can have completely different effects on each variable, a case considered in figure 1.3(e). There are five variables V_1 to V_5, and each one has a different behaviour in response to a common set of external factors. In the example, V_1 is positively affected by factors of type 1 and negatively by type 2; variable V_2 shows the opposite behaviour; variable V_3 shows a more complicated behaviour, with the same reactions as variable V_2, but with internal reactions to the behaviour of variable V_1. Response times also vary: V_1 responds faster than V_2. Variable V_4 is unaffected by external factors of type 1, but shows a negative reaction to type 2. Finally, V_5 is affected negatively by all factors, until it reaches a low point, from where it rises suddenly. This latter type of variable has been explored in threshold analysis and catastrophe theory. Threshold analysis (Malisz, 1969) explores several types of responses, stating that some variables, particularly those related to costs, show important discontinuities at particular points.

The simulation of the dynamic behaviour of variables can be carried out in a continuous form or by using discrete time intervals. Discrete dynamics is usually preferred, particularly because time-series data are available only at discrete intervals. Within the discrete approach there are two main types of model that can be constructed: with endogenous and with exogenous dynamics. These two alternatives are presented in figure 1.4.

Figure 1.4(a) shows a system with endogenous dynamics, where the set of variables changes as a function of its previous state at $t - 1$. Naturally, some exogenous inputs are required to tell the model how variables change from one period to the next; these are normally rates of change calculated from past data. In figure 1.4(b) a case of exogenous dynamics is shown, where the system is supposed to change only as a result of changes in the exogenous variables. In an urban or regional system, both types of dynamics will be found simultaneously, as shown in figure 1.4(c). In this case, there are exogenous variables that change through time and affect the system in each period, and there are also endogenous variables that are a function of the previous time period. Note that the possibility of the system state at each time period affecting the external variables of the next has also been considered (diagonal arrows).

Exogenous factors are elements outside the system under consideration, which can either be elements of the system at a higher level of resolution (e.g. a region for a city), or elements at the same level that have been left out of the system definition (e.g. complex socio-economic

Figure 1.4. (a) Endogenous, (b) exogenous, and (c) mixed system dynamics



or cultural elements). The endogenous factors refer to dynamic elements that occur within the system as defined.

However, exogenous and endogenous factors alike will produce effects in the system with different response times, and it can be said in general that the speed with which they react will depend on the speed of the actors involved. For instance, a car driver going from *A* to *B* makes a decision every day about which route he should choose as the most convenient. If a road is suddenly closed or if a new facility becomes available, he will probably change the characteristics of his trip almost immediately. In this case, time lags will be very short. If, though, the question is about migrating from one city to another, the response time might be considerably longer.

1.7 Conclusions

The emphasis of this chapter has been on questions of epistemology and methods. It began by arguing that in the last few decades methods in social and natural sciences have come considerably closer. This was called the process of convergency. With this argument in mind, the scientific method as conceived by Popper was presented, and later adapted to the social sciences. It was argued that Popper's proposition is not only applicable to social sciences, but if properly adapted, it can be applied to social processes in general. It can be used to explain the way people make everyday decisions and how they interact within the social structure. Two special cases were discussed in some detail: the social scientist and the planner.

It was also argued that one of the key elements in the resulting method for the social sciences is the use of models, because they enable the analyst to perform experiments, which would be otherwise impossible. This explains the growing importance of models, not only as a tool for policy testing in planning, but also as a way of testing theories in social sciences.

2

Micro-economic theories of land use

The earliest models which sought to explain the spatial distribution of activities were based on micro-economic theory. They took as their central concern two related questions: what is the rationale that regulates the location of activities, and how does land rent or land values emerge from this process? The models which evolved from this micro-economic approach started by adopting the conceptual framework of classical and neo-classical economics.

In this approach, the location of activities is seen as the outcome of a combined market mechanism involving three basic elements: commodities, land and transport. On the one hand, land is required to perform productive operations, or for the satisfaction of residential needs. On the other hand, transport is required in order to move surplus production or labour. A farmer, for instance, wants to sell his surplus in the market-place, and the resident of a city wants to sell his labour where there is a demand for it. The process of exchange involves transportation, either of commodities to the market, or of residents to their place of work.

Land is considered to be a large featureless plain and infinitely available, so that in principle there would be no need to pay for it. What gives land a differential quality is the cost of travel or accessibility, which is the main factor in the generation of land values.

Micro-economic theories of land use look at the process of activity location and rent from the point of view of the individual resident or firm. Activities will compete for the consumption of land, and once equilibrium has been reached, they will have chosen a site of a particular size, such that the cost of land and transport they have to pay optimises their utility. Competition will also determine the price of land in each location and the price of commodities at the market-place.

The distance between agricultural producers and the market-place is taken as an example in Von Thünen's (1826) theory of location, which explains the way in which transport costs affect agricultural producers and consumers, the consequences for the process of land allocation and rent, and the price of agricultural commodities. Transport is an essential element used by Weber (1909) to explain the location of industry with respect to a market of industrial commodities, although in this case the consumption of land is not made explicit. Christaller (1933) and Lösch

(1940) also used a similar concept to explain the formation of *market areas* and their geometrical arrangement to form regions. The urban case is explained by Wingo (1961) and Alonso (1964) as one in which residents and firms compete for the use of land, generating an equilibrium pattern of location and rents around a single centre of employment. More elaborate versions of these models are given by Muth (1968) and Mills (1969).

In this chapter, some of this work will be reviewed. Von Thünen's model is taken as a basis to arrive at the bare bones of spatial micro-economic theory, by extending the original model in several ways. After this, the models of Wingo and Alonso are described, but it will be argued that an extended version of Von Thünen's model can be as general as the more modern versions. A computer program is used to illustrate the theory. The explanation ends with the models of Christaller and Lösch, to show how regions of multiple centres can be formed.

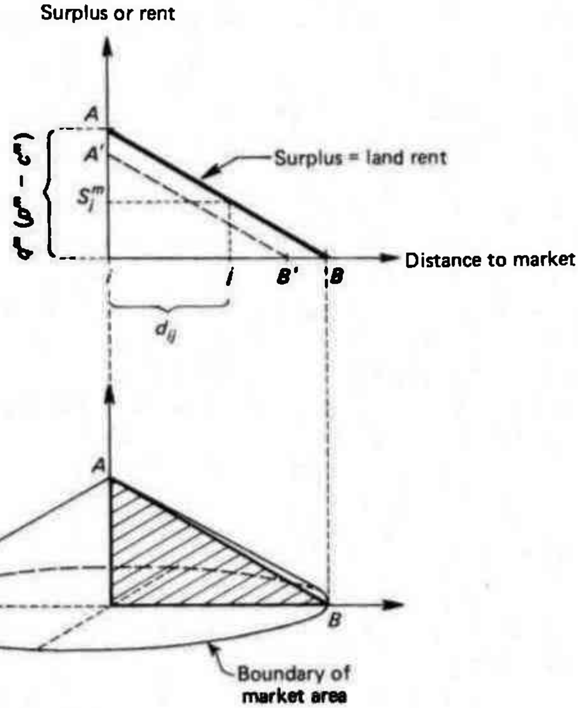
2.1 Von Thünen's isolated state

Von Thünen made the first attempt to explain the effect of transport costs on the location of activities and the functioning of the land market. For this purpose he used as an example an idealised agricultural region, at the centre of which there is a single market-place where a large number of producers want to sell their products. Land is in the hands of a large number of landowners who are willing to rent their properties to the highest bidder, that is, to the producer who is willing to pay a highest rent. A number of other simplifying assumptions are included in Von Thünen's model:

- the system under consideration is *closed* (isolated state) in the sense that there is no interaction with other regions, and that once equilibrium has been reached, no actors leave or enter the system;
- land is homogeneous in terms of fertility, productivity and transport costs, i.e. the cost of transport per unit of distance is constant in all directions;
- there is only one market centre where all agricultural commodities are sold;
- there is a large number of producers trying to maximise benefits, and a large number of landlords trying to maximise rent; hence, neither can individually control prices;
- there is no cost involved when a producer or landlord decides to enter or leave the system.

Having made these assumptions, Von Thünen proceeds to analyse the conditions of the land market for each individual producer. Consider the example of figure 2.1 where there is a single market-place at point i. Consider, next, the conditions in a particular point j at a distance d_{ij} from the market-place. The surplus to the producer of the single commodity m at point j, S_j^m, will depend on the amount produced and on the

Figure 2.1. Von Thünen's single commodity model



price obtained in the market, that is:

$$S_j^m = q^m(p^m - c^m - k^m d_{ij}) \tag{2.1}$$

where q^m is the amount of commodity m produced per unit of land; p^m is the price per unit of commodity m at the market-place; c^m is the cost of production of one unit of commodity m; k^m is the cost of transport of one unit of commodity m per unit of distance; d_{ij} is the distance from j to the market-place.

Equation (2.1) is a linear function (line AB in figure 2.1), decaying from a maximum (point A) when the distance to the market is zero, that is, when $S_j^m = q^m(p^m - c^m)$, to a minimum (point B), where $S_i^m = 0$, that is, when the cost of transport is equal to $p_m - c_m$. Beyond this point, the cost of production plus the cost of transport exceeds the price in the market, so that producers will leave the market.

Landowners, on the other hand, will rent their land to the highest bidder. In a given location j, producers will compete to secure land for themselves. Each producer will try to outbid the others by increasing the amount of money he is willing to pay for the land, but only up to the maximum he can afford, which is S_j^m. This is roughly what Alonso (1964) later called *bid price*. Landowners cannot increase rent above this value because producers would rather leave than operte at a loss. Thus, the resulting land rent profile in the region (line AB) exactly matches the profile of surplus. If line AB is rotated around its axis, a circle with centre i will be described which represents the area where commodity m is produced.

Figure 2.1 also shows the effect on surplus and land rent of a fall in the price of commodity *m*. As a result, the surplus or land rent curve *AB* will shift to *A'B'*, *AA'* representing the drop in price.

Up to now only one commodity has been considered. If producers of other commodities are introduced into the system, landowners will rent the land to the producer of the commodity that renders the highest surplus, since they are the ones that will be willing to pay the highest rent. Figure 2.2 shows the surplus profiles of three commodities: $m = 1, 2, 3$. Each one has a different slope depending on the relation between transport costs and yield. Curve *EF*, for instance, is steeper than *AB*, because the transport cost per unit of distance of commodity 2 is higher than that of commodity 1. The price of commodity 2 in the market is higher than that of commodity 1, hence it intersects the abscissa at a higher point, point *E* instead of *A*.

The producers of these commodities will compete for land. At a location very near the centre, say point *k*, landowners can choose between producers of commodities 1, 2 or 3. Because the surplus of commodity 2 is the highest at this point, they will assign their land to a producer of 2, and will receive a rent equal to S_k^2. As we move outwards, commodity 2 will still produce a surplus higher than the others, and consequently landowners will rent their land to producers of 2, until a point is reached where its surplus profile intersects that of commodity 1. All the land up to this point of intersection will be allocated to producers of commodity 2, so that they will occupy a circle of land around the market-place. Beyond the point of intersection,

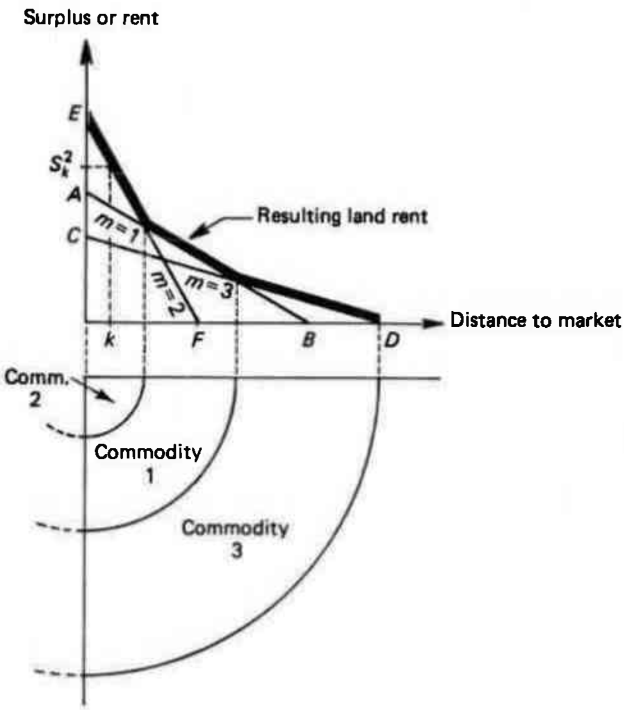


Figure 2.2. Von Thünen's model with several commodities

producers of commodity 1 will be willing to pay more for the land than any other, so that land will be assigned to them, until profile 1 intersects profile 3. From there onwards producers of 3 will become the highest bidders until the margin is reached. The resulting land rent is represented by the dribble profile *ED*, and land uses will form concentric rings around the market centre.

As can be seen, for each commodity there will be a location *at the margin*, where its production stops, which is, at the same time, the location *nearest to the centre* of the next commodity along the chain. Land value at the margin for a particular commodity will equal land value nearest to the centre for the next commodity.

2.2 Calculating Von Thünen's basic model

The following data is required to calculate the model:

M number of commodities
q^m amount of land required to produce one unit of commodity m
k^m cost of transport per unit of distance for commodity m
c^m cost of production for commodity m
D^m total demand for commodity m in the market-place

With these exogenous inputs, the model must calculate land rent at every point, and the location at the margin and price for each commodity. The calculation proceeds in the following steps:

(a) Calculate the slope of the surplus/rent profile of each commodity, which is, from equation (2.1):

$$s^m = k^m/q^m \tag{2.2}$$

(b) Sort commodities in order of slope, such that $s^1 > s^2 > s^3 > \ldots > s^M$.

(c) Calculate the area required for the production of each commodity, A^m, as the demand for product m multiplied by the amount of land required to produce one unit:

$$A^m = D^m q^m \tag{2.3}$$

(d) Calculate the distance at the margin for each commodity. Since commodities are arranged as concentric rings around the market, the radius of the first commodity, i.e. the one with the steepest surplus/rent profile, can be calculated from the standard equation of the area of a circle. The second distance will be the radius of a circle with an area equal to that for the first plus the second commodities, and so on. For any commodity m, the distance at the margin will be:

$$d_i^m = \sqrt{\left[\left(\sum_{n=1}^{m} A^n\right)\Big/\pi\right]} \tag{2.4}$$

(e) Calculate land rent at the margin and the price for each commodity. For the last commodity, land rent will be zero, and the price will be:

$$p^M = k^M d_i^M + c^M \tag{2.5}$$

This will be the starting point of the calculation, which will proceed from the last commodity towards the centre. For any commodity m, rent at the margin, $\bar{S}^m$, will be:

$$\bar{S}^m = \frac{p^{m+1} - c^{m+1} - k^{m+1} d_i^{m+1}}{q^{m+1}} \tag{2.6}$$

and the price will be:

$$p^m = \bar{S}^m q^m + c^m + k^m d_i^m \tag{2.7}$$

Finally rent at the centre will be:

$$S_0 = \frac{p^1 - c^1}{q^1} \tag{2.8}$$

Figure 2.3 shows three examples (calculated with program VONTH), corresponding to one, two and three commodities in sequence, showing inputs and outputs. Example (a) has only one commodity, oats; the resulting values such as steepness, area, distance to the market and land value, can be easily checked from the inputs. Example (b) assumes that a second commodity joins the system: barley. The introduction of this new commodity affects the previous system in several ways. Because the relationship between transport cost and yield (equation (2.2)) of the new commodity is smaller than that of oats, producers of barley will locate in an outer ring. The cultivated area will extend from a previous radius of 12.49 units to 16.25 to accommodate the new product. Land rent will increase in all locations, and the price of oats will rise from its previous 862 to 993. The difference in price will be transferred to landowners in the form of rent.

Figure 2.3(c) shows the result of introducing a third commodity: wheat. Since this new commodity has the highest transport cost/yield relationship, the effect will be significant: wheat will now take over the land previously occupied by producers of oats and some of the producers of barley, pushing these outwards. Land values will increase considerably in all locations, and so will the price of the old commodities. Total cultivated area will also increase.

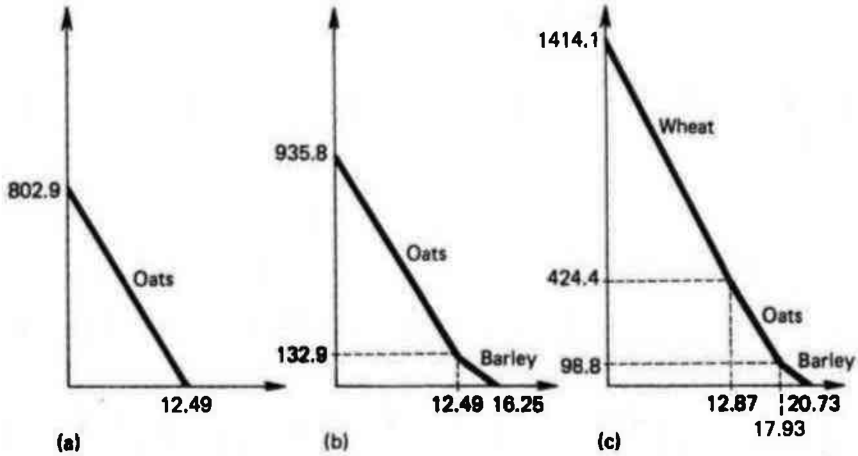
This example broadly parallels the process of evolution in socio-spatial systems. As a society develops, more and more commodities enter the system. As a result, the size of the market area increases, there is more demand and production, and land values increase.

2.3 Extending Von Thünen's model through elasticities

2.3.1 Demand elasticities

In the previous example, demand was fixed and a given input to the

Inputs	(a)	(b)	(c)
No. of commodities	1	2	3
Added commodity	Oats	Barley	Wheat
Yield	0.700	0.850	0.650
Transport cost per unit distance	45.0	30.0	50.0
Demand	700.0	400.0	800.0
Cost of production per unit	300.0	500.0	250.0



Outputs	(a)	(b)	(c)
Steepness of surplus/rent curve			
Oats	64.29	64.29	64.29
Barley	–	35.29	35.29
Wheat	–	–	76.92
Cultivated area			
Oats	490.0	490.0	490.0
Barley	–	340.0	340.0
Wheat	–	–	520.0
Total	490.0	830.0	1350.0
Distance at the margin			
Oats	12.49	12.49	17.93
Barley	–	16.25	20.73
Wheat	–	–	12.87
Rent at the margin			
Oats	0.0	132.9	98.8
Barley	–	0.0	0.0
Wheat	–	–	424.4
Rent nearest to the market	802.9	935.8	1414.1
Price of commodities			
Oats	826.0	955.0	1176.0
Barley	–	987.6	1121.9
Wheat	–	–	1169.1

Figure 2.3. A numerical example of Von Thünen's model with (a) one, (b) two, and (c) three commodities in sequence

model. As more commodities entered the system, prices increased, but demand remained the same. A more realistic assumption is that, as prices increase, demand is reduced. This phenomenon, known as *elasticity*, is of paramount importance in economic and spatial systems.

Fixed demand, then, must be replaced by a *demand function*, such as the one shown in figure 2.4. Since the amount consumed will drop as price increases, demand functions will have a negative slope. The speed at which demand is reduced, the elasticity of demand, will vary according to the type of commodity and type of consumer. In many cases, for instance, low income populations have high elasticities with respect to most commodities. If the consumption of a particular commodity is considered essential, elasticity will be low. In the example of the previous section, demand for all commodities was assumed to be completely inelastic.

Mathematically, a demand function can be represented in many ways, provided it is a decaying function. A linear function with a negative slope would be the simplest choice. However, a function with a constant slope is unreal, because of the principle of *diminishing returns*, which states that the level of satisfaction derived from additional units of consumption tends to decrease. This means that a point is reached where, even if the price of a particular commodity nears zero, the amount consumed will not exceed an upper limit. As the price increases, demand will drop rapidly at the beginning, more slowly afterwards, and will be asymptotic to zero at a certain point, depending on the consumer and the commodity.

The most commonly used functions with a varying negative slope are the power and exponential functions. A decaying power function is conveniently asymptotic to zero at some point, but does not have an upper limit. This means that, as the price of a commodity tends to zero,

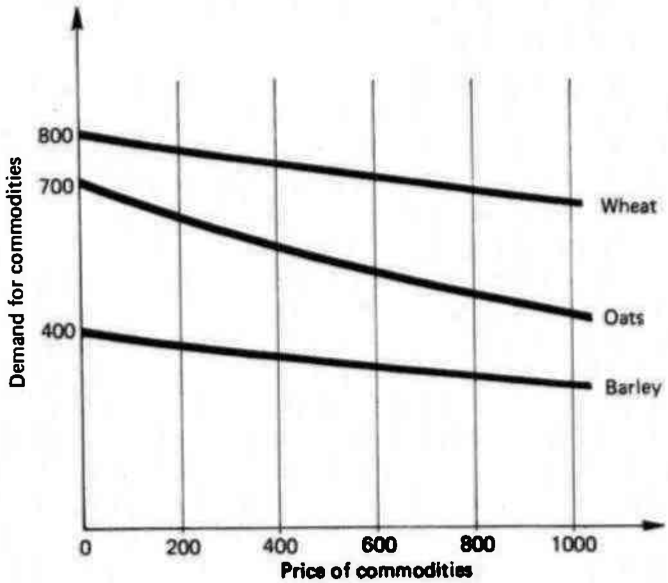


Figure 2.4. Demand functions for commodities to extend Von Thünen's model

the amount consumed would tend to infinity. A negative exponential function is also asymptotic to zero at some point, but does have an upper limit, making it the natural choice.

Going back to the example above, exogenously given demands must be replaced by demand functions of the form:

$$D^m = \sigma^m \exp(-\tau^m p^m) \tag{2.9}$$

where D^m is the resulting demand for commodity m given that p^m is the price, and where σ^m is a given parameter representing the maximum amount of commodity m consumed if the price is zero, and τ^m is the elasticity parameter. In equation (2.9), if $p^m = 0$, then $\exp(-\tau^m p^m) = 1$, so that $D^m = \sigma^m$. At the other extreme, if p^m is large, $\exp(-\tau^m p^m)$ tends to zero, and so will D^m. If parameter τ^m is large, D^m will drop quickly as prices increase. On the other hand if $\tau^m = 0$, then D^m will be constant and equal to σ^m (a horizontal line), representing the absolutely inelastic case. The example of the previous section would be achieved by setting all τ^m's to zero in (2.9).

In order to calculate the model, apart from changing given demand by demand parameters σ^m and τ^m, the process would have to be modified as follows:

(a) Set the prices of all commodities to zero. This is necessary because at the beginning of the calculation, prices are not yet known.
(b) Calculate the demand for each commodity, applying equation (2.9).
(c) Calculate steps (a) to (e) of the previous section. This will determine the slope of the surplus profile for each commodity and consequently the order in which they will be located around the market. It will also provide the area, the distance and rent at the margin, and the price of each commodity.
(d) Evaluate convergency. If achieved, the calculation stops. Otherwise, return to step (b).

In step (b) demand was calculated as a function of prices, which were initially set to zero ($D^m = \sigma^m$). At the end of step (c), a new set of prices is calculated, which will affect demand. The calculation process must then return to step (b), where demand must be re-calculated, which in turn will produce a new set of prices. The calculation process becomes *iterative*, repeating the same sequence of calculations again and again.

As more iterations are performed, prices will rise and demand will fall continuously, but because of the shape of demand curves with diminishing negative slopes, a point will be reached where all variables in the system are changing very little with respect to the previous iteration. This is where calculation should stop; it is called *convergency*.

In order to measure convergency, one of the variables must be chosen as indicator, preferably a variable which is affected by all the others. Total land consumed is the obvious choice, since producers of all commodities will consume land. The percentage difference between land consumption in one iteration and its preceding one becomes the convergency indicator, and must be compared with a given criterion for convergence.

Figure 2.5 shows the same example of figure 2.3 after the introduction of demand elasticities, calculated with program VONTH. Previous demands are taken here as maxima (σ''') and elasticity parameters are added as input data (τ'''). The demand functions plotted in figure 2.4 correspond, in fact, to the values adopted here.

The main difference in this example with respect to the previous one is that demand figures are now endogenous and are allowed to drop as prices increase. For example, in figure 2.3(a) demand for oats was fixed at 700, and here, because of elasticity, demand drops to 480 after convergence. When barley is introduced in figure 2.5(b), demand for oats drops further to 462, and to 419 when wheat is introduced. All other variables change accordingly, except the steepness of the surplus/rent profiles, which depend only on transport costs and yield. The ordering of the commodities around the market thus remains unchanged. Areas and distances to the market are reduced because fewer units of each commodity must be produced. Land values will also decrease as consumers adjust their demand to prices. All cases converged after three iterations to meet a criterion for convergence of one per cent.

The main conclusion to be added when demand elasticities are introduced is that, as more commodities join the system, people consume less of the previous goods in order to consume more of the new ones (*substitutions*). The degree to which this happens is determined by the elasticity parameters, a relation that is called *cross-elasticities* in economics.

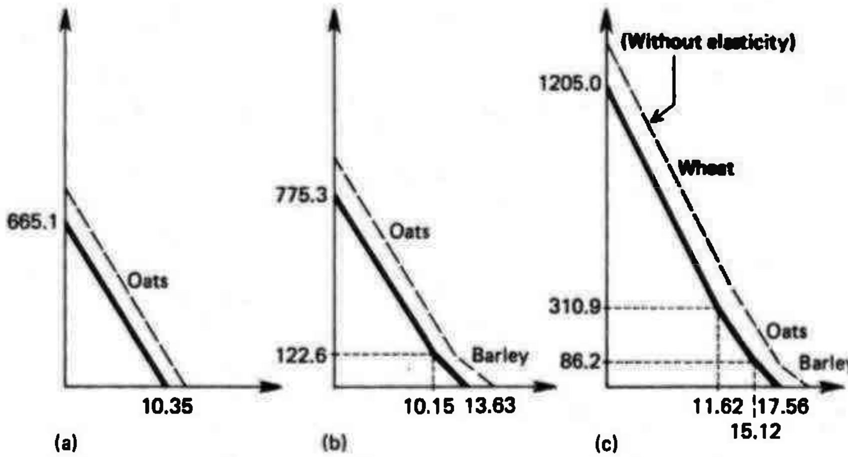
From a mathematical point of view, the introduction of elasticities implies an iterative calculation process because the system ceases to be linear. Iterative algorithms such as the one described above can be very powerful in solving these non-linear systems in a simple and computationally efficient way. For this reason, in the rest of this book, iterative algorithms will be used extensively.

2.3.2 Land consumption elasticities

In the previous examples, the amount of land required to produce one unit of each commodity was fixed and exogenously given to the model. A more realistic assumption is that, as the price of land rises, producers are induced to invest more in machinery and fertilisers in order to produce more per unit of land. Producers of some commodities will be able to do so more than others, as shown in figure 2.6, for technological reasons.

This means that now, instead of given yields per commodity, land demand functions can be introduced. Land required to produce one unit of a particular commodity will be maximum when land rent is zero, and will decay as land values increase. In this case, however, it is convenient to introduce a minimum, below which no extra amounts of machinery and fertilisers will achieve an increase in productivity. If a negative

Inputs	(a)	(b)	(c)
No. of commodities	1	2	3
Added commodity	Oats	Barley	Wheat
Yield	0.700	0.850	0.650
Transport cost per unit distance	45.0	30.0	50.0
Maximum demand (σ^m)	700.0	400.0	800.0
Demand elasticity (τ^m)	0.0005	0.0003	0.0002
Cost of production per unit	300.0	500.0	250.0



Outputs	(a)	(b)	(c)
Steepness of surplus/rent curve			
Oats	64.29	64.29	64.29
Barley	–	35.29	35.29
Wheat	–	–	76.92
Cultivated area			
Oats	336.0	324.0	294.0
Barley	–	260.0	251.0
Wheat	–	–	424.0
Total	336.0	584.0	969.0
Distance at the margin			
Oats	10.35	10.15	15.12
Barley	–	13.63	17.56
Wheat	–	–	11.62
Rent at the margin			
Oats	0.0	122.6	86.2
Barley	–	0.0	0.0
Wheat	–	–	310.9
Rent nearest to the market	665.1	775.3	1205.0
Price of commodities			
Oats	765.6	842.7	1040.7
Barley	–	908.8	1026.8
Wheat	–	–	1033.2
Demand for commodities			
Oats	480.4	462.6	419.6
Barley	–	305.4	295.0
Wheat	–	–	652.9

Figure 2.5. A numerical example of Von Thünen's model with (a) one, (b) two, and (c) three commodities in sequence and elasticities in the consumption of commodities

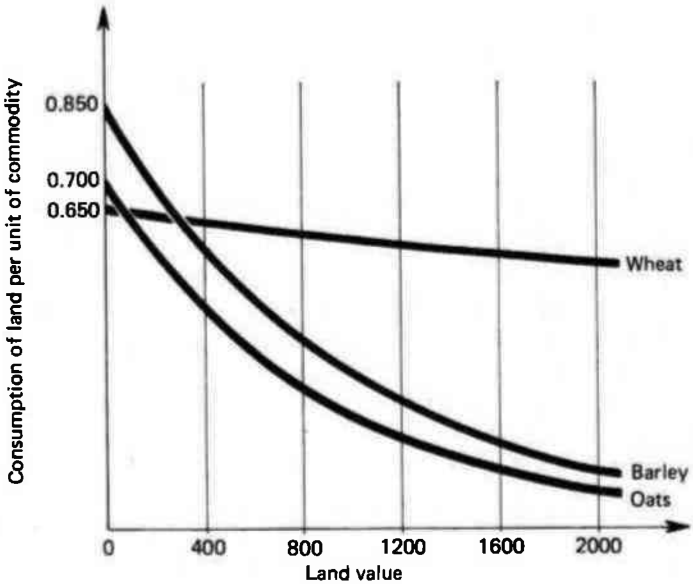


Figure 2.6. Demand functions for land to extend Von Thünen's model

exponential function is adopted, the land demand function would be:

$$L^m = l^m \exp(-\delta^m \bar{S}^m), \quad \text{subject to } L^m \leqslant L^{m*} \tag{2.10}$$

where L^m is the amount of land required to produce one unit of commodity m; l^m is the maximum amount of land consumed when the land rent is zero; δ^m is the elasticity parameter; $\bar{S}^m$ is the average land value for commodity m; L^{m*} is the minimum amount of land required to produce one unit of commodity m. Note that equation (2.6) determines surplus/land values at the margin for each commodity, and that surplus/land value at the point nearest to the market is the value at the margin of the preceding one. In order to apply equation (2.10), the average $\bar{S}^m$ must be estimated. Here $(S^m + S^{m+1})/2$ is used as a rough approximation.

In order to include land consumption elasticities, the calculation process has to be modified as follows:

(a) Set prices of all commodities to zero. Set land values to zero for all commodities.
(b) Calculate demand for each commodity applying equation (2.9), and yield applying equation (2.10).
(c) Calculate the slope of the surplus/land value profile for each commodity, and consequently the order in which they will be located around the market. Calculate the area, distance, land values and price for each commodity.
(d) Evaluate convergence. If achieved, calculation stops. Otherwise go back to step (b).

In step (b) both demand and yield are calculated as a function of prices and land values, initially set to zero. At the end of step (d) a new set of prices and land values are calculated, which will affect demand and yield. The calculation process must then return to (b) to re-estimate demand and yield and then proceed with all other calculations in (c). As in the previous case, land consumption can be chosen to evaluate convergence in each iteration.

Figure 2.7 shows the results of adding land consumption elasticities to the previous example of figure 2.5. Previous yield figures are taken as maximum land consumptions (l^m), adding elasticities (δ^m) as well as minimum consumptions (L^{m*}). When the first commodity is introduced in figure 2.7(a), yield previously fixed at 0.70 is now allowed to drop with an elasticity of 0.0012 towards a minimum of 0.20. As land rent increases, producers of oats increase the productivity of land at the same time as consumers reduce their demand. The result is a steeper surplus/ land value profile, less and more-expensive land, cheaper oats and consequently more demand. When barley is introduced, similar effects take place, but when wheat enters the system with a very small capacity to increase productivity, a somewhat dramatic effect takes place. Because producers of oats are far more capable of increasing productivity, the order in which commodities arrange themselves around the market is reversed. In figure 2.5(c) the order was wheat–oats–barley, and in figure 2.7(c) the order becomes oats–wheat–barley. Comparing both sets of results, it is easy to see that areas are smaller, land values are higher and that there is more demand for cheaper commodities.

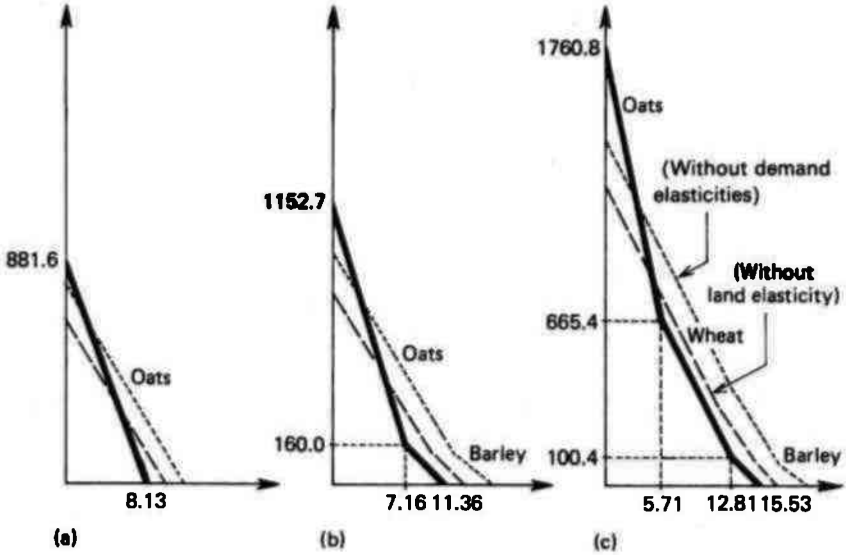
In general, it can be concluded that when yield is allowed to vary, land values seem to be the main force behind technical innovation. As more and more commodities enter the system, land values rise, pushing the price of commodities up, and consequently reducing demand. Producers are then forced to increase productivity to recover demand, making the whole system more efficient. This also explains why a less-developed society with fewer commodities will tend to show lower levels of productivity in each sector.

2.4 Wingo's transportation and land use model

The application of the model of Von Thünen to the urban case follows a natural path. The model developed by Wingo (1961) deals with the urban phenomenon, and in particular with the relationship between transport cost, the location of activities, and land values. Wingo

Inputs	(a)	(b)	(c)
No. of commodities	1	2	3
Added commodity	Oats	Barley	Wheat
Maximum land consumption (l^m)	0.700	0.850	0.650
Land consumption elasticity (δ^m)	0.0012	0.0010	0.0001
Transport cost per unit distance	45.0	30.0	50.0
Maximum demand (σ^m)	700.0	400.0	800.0
Demand elasticity (τ^m)	0.0005	0.0003	0.0002
Cost of production per unit	300.0	500.0	250.0

Figure 2.7. (continued overleaf)



Outputs (after convergence)	(a)	(b)	(c)
Steepness of surplus/rent curve			
Oats	108.50	138.55	191.97
Barley	–	38.18	36.96
Wheat	–	–	79.48
Cultivated area			
Oats	207.0	161.0	102.0
Barley	–	244.0	242.0
Wheat	–	–	414.0
Total	207.0	405.0	758.0
Distance at the margin			
Oats	8.13	7.16	5.71
Barley	–	11.36	15.53
Wheat	–	–	12.81
Rent at the margin			
Oats	0.0	160.0	665.4
Barley	–	0.0	0.0
Wheat	–	–	100.4
Rent nearest to the market	881.6	1152.7	1760.8
Demand for commodities			
Oats	500.1	496.5	436.4
Barley	–	310.4	298.1
Wheat	–	–	657.5
Yield			
Oats	0.41	0.32	0.23
Barley	–	0.79	0.81
Wheat	–	–	0.63
Price of commodities			
Oats	665.6	674.4	712.8
Barley	–	840.7	966.0
Wheat	–	–	953.9

Figure 2.7. A numerical example of Von Thünen's model with (a) one, (b) two, and (c) three commodities in sequence and elasticities in the consumption of land

proposes the principle of *complementarity* between the cost of transport and land rent; this means that whatever the location of a household, a fixed amount of money will be spent on land and transport. In the urban case, the introduction of a land consumption function becomes obligatory, because residential activities have an elasticity too high to be ignored.

As in Von Thünen's model, a number of simplifying assumptions are made by Wingo. In the idealised urban area there is only one centre of employment where all residents work. All residents have the same income and preferences. Prices of all goods and services are equal in every location, except for land. Land has a positive utility, that is, households will always try to consume more of it if possible. Transport costs are a linear function of the distance to the single employment centre.

Let i denote the location of the employment centre, j a typical location of a household, and m the location at the margin where the urban area ends. c_{ij} represents the transport cost from any location j to the employment centre. Rent at any point, R_j, is simply defined as the price per unit of land, times the quantity consumed, l_j:

$$R_j = r_j l_j \tag{2.11}$$

The hypothesis of complementarity between land rent and transport can be expressed as:

$$K = R_j + c_{ij} \tag{2.12}$$

where K is the constant amount that households spend on transport and land. At the centre itself, transport cost is zero ($c_{ii} = 0$), and thus rent is maximum ($R_i = K$). At the margin, transport cost is maximum ($c_{im} = K$), and thus rent is minimum ($R_m = 0$). As in Von Thünen, any savings in transport costs with respect to costs at the margin will be transferred to land in the form of rent. Land rent at intermediate points is then the cost of transport not incurred:

$$R_j = c_{im} - c_{ij} \tag{2.13}$$

Since land has a positive utility, a demand function can be established in which, as the price of land increases, less land is consumed by a household. Wingo proposes a power function, but an exponential function can be adopted advantageously such as equation (2.10), which adapted to this case becomes:

$$L_j = l \exp(-\delta r_j) \tag{2.14}$$

where L_j is the amount of land consumed by a household located in j, l is the maximum amount of land households of the single socio-economic group are willing to consume, and δ is the elasticity parameter. At any point j, equations (2.11) and (2.13) determine the amount of money to be spent on rent, and (2.14) determines the amount of land (plot size) to be consumed. Residential density would be the inverse of plot site, i.e. $(L_j)^{-1}$. Assuming that land is homogeneous and that the region has a circular shape, the amount of land available at a distance d_{ij} from the centre, A_j, will be $\pi(d_{ij})^2$, and the total size of the urban area will be $A^* = \pi(d_{im})^2$. Hence, a market equilibrium condition can be established, whereby total population P must be equal to the integral of the density over the whole area of the city:

$$P = 2\pi \int_0^{d_{im}} d_{ij}(L_j)^{-1}\,\mathrm{d}\,d_{ij} \tag{2.15}$$

The model in its simplest form can be represented graphically as in figure 2.8. In diagram (a), point m represents the margin where households spend all their budget K on transport. It is assumed that all households have wages made up of a basic amount W plus the amount K to cover the combined cost of land and travel. At any point j households spend an amount c_{ij} on transport, and are left with a surplus $K - c_{ij}$ which will be transferred to land in the form of rent R_j. In figure 2.8(b) both the rent per unit of land curve and the corresponding residential density curve are plotted together (density is upside-down). Rent per unit of land decays from a maximum K to zero at the margin. Density will also fall from l^{-1} at the centre to zero at the margin.

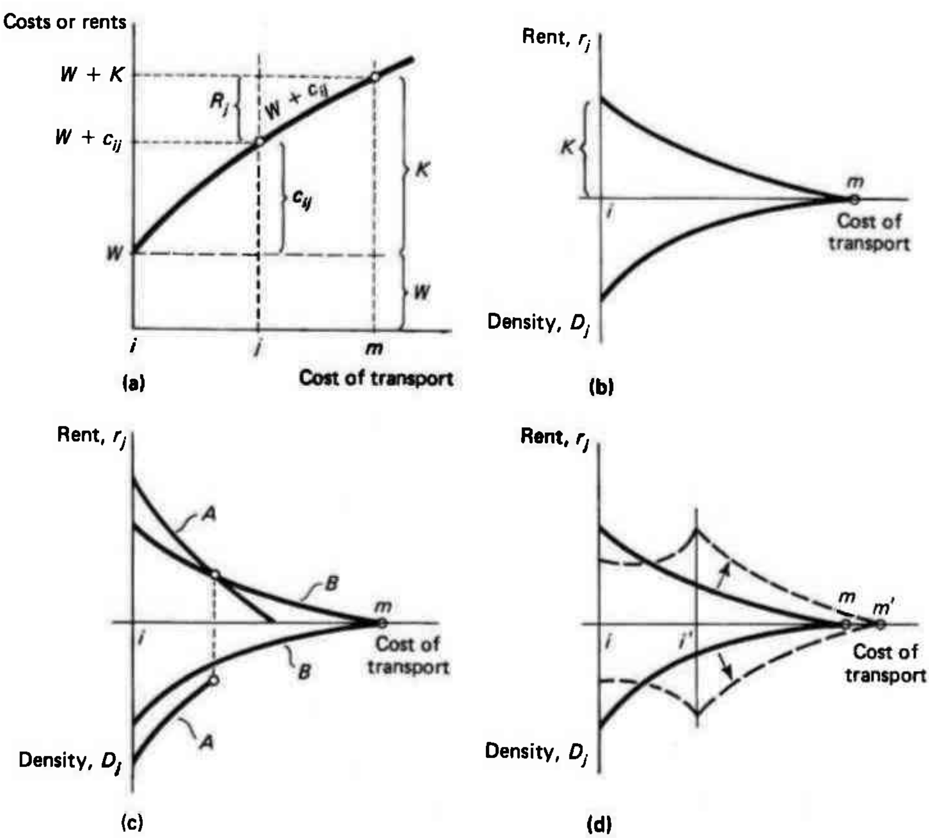


Figure 2.8. Graphical representation of some aspects of Wingo's model

Wingo then drops some of the simplifying assumptions. Although an analytical solution will no longer be possible, the effects can be described graphically. Figure 2.8(c) considers two types of households corresponding to two different income groups. Curve A represents a type of household particularly sensitive to transport costs, but not so much to the cost of land; households of this type will tend to locate near the centre, consuming a small amount of land. Curve B represents households with opposite preferences. Group A, then, has a steeper curve both for rent and density. With such extreme groups, discontinuities can appear in the density curve, but as more intermediate groups are added, this step-wise effect should smooth out. Finally figure 2.8(d) shows the effect of introducing a second employment centre at point i'. The margin will shift from m to m', and both rent and density curves will show two peaks in proportion to the magnitude of the employment located in each centre.

2.5 Alonso's location and land use model

The model of Alonso has many similarities to that of Wingo, although both were developed independently at approximately the same time. Both are urban models and share the same simplifying assumptions, such as homogeneous land, a single employment centre, and elastic demand for land. In the case of Alonso, though, users of land can be either households of different income groups or firms, and the budget constraint equation considers three elements instead of two: land rent, transportation expenditure, and a composite expenditure representing all other goods and services consumed by the household or firm. The constant amount is then equal to total income:

$$y = l_j r_j + c_{ij} + p_z z \tag{2.16}$$

where y is the income, p_z is the price of the composite good, and z is the number of units of the composite good consumed.

The three elements of equation (2.16) can be combined in many different ways. An individual household can consume, say, less transport in order to consume more land and composite good. For each household there will be an infinite number of possible combinations of these three elements, giving rise to a three-dimensional surface called the *locus of opportunity surface*, as shown in figure 2.9. Each point in the surface will render the household a particular level of satisfaction or *utility*. Users will move along their locus of opportunity surface in order to maximise their utility.

In figure 2.9 three typical sections across the locus of opportunity surface are shown. In the second one, the amount of land that can be consumed in each location is plotted against the consumption of the composite good z, keeping transport cost at a constant level. This results in a linear function, because as more is spent on land, less is available for the consumption of composite good z.

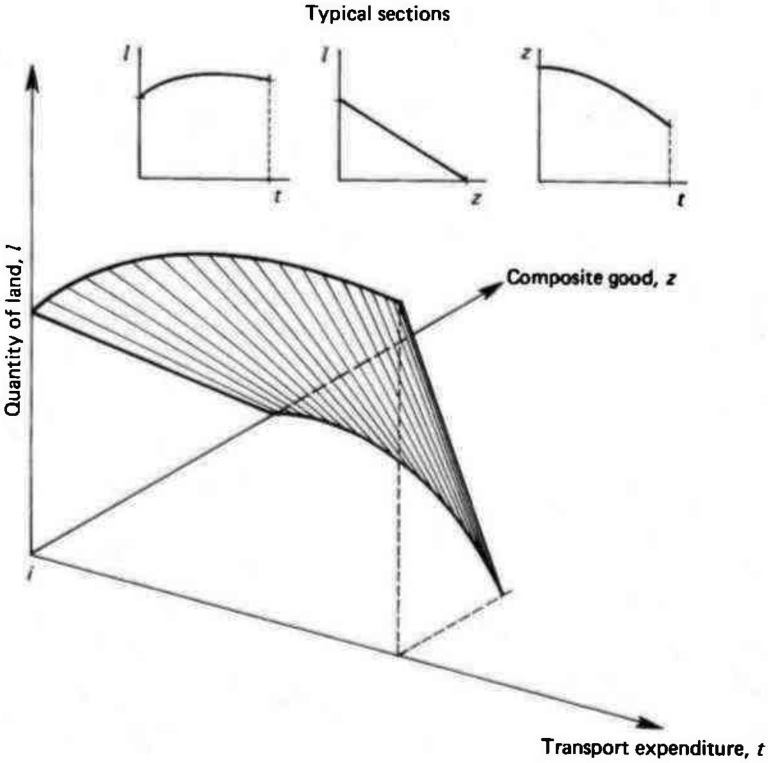


Figure 2.9. Locus of opportunity surface in Alonso's model

The first diagram shows the relation between the amount of land consumed and transport expenditure, when z is kept constant. As an individual user moves away from the centre, he can consume more of an increasingly cheaper land. However, a point is reached where increasing transport costs compensate savings in land, and from there onwards users are forced to reduce their consumption of land. A similar situation is shown in the third diagram, where the relation between transport expenditure and the consumption of the composite good is shown, assuming that a constant amount of money is spent on land.

If each combination of land, transport and composite good renders a certain level of utility, sections of an *indifference map* can be drawn, as in figure 2.10. In the first diagram, land is plotted against transport, assuming a constant consumption of z. The different points along the curve show combinations of land and transport that produce the same utility to individuals, so that they will be indifferent or will show equal preferences as to which combination they will adopt. The odd shape is produced by the different character of the consumption of land and transport: while land has a positive utility, transport has a negative utility. In other words, users will try to consume as much land as possible and as little transport as possible. The curve shows that, as the consumption of transport increases, the consumption of land increases also in order to maintain the same level of utility. Individuals of the

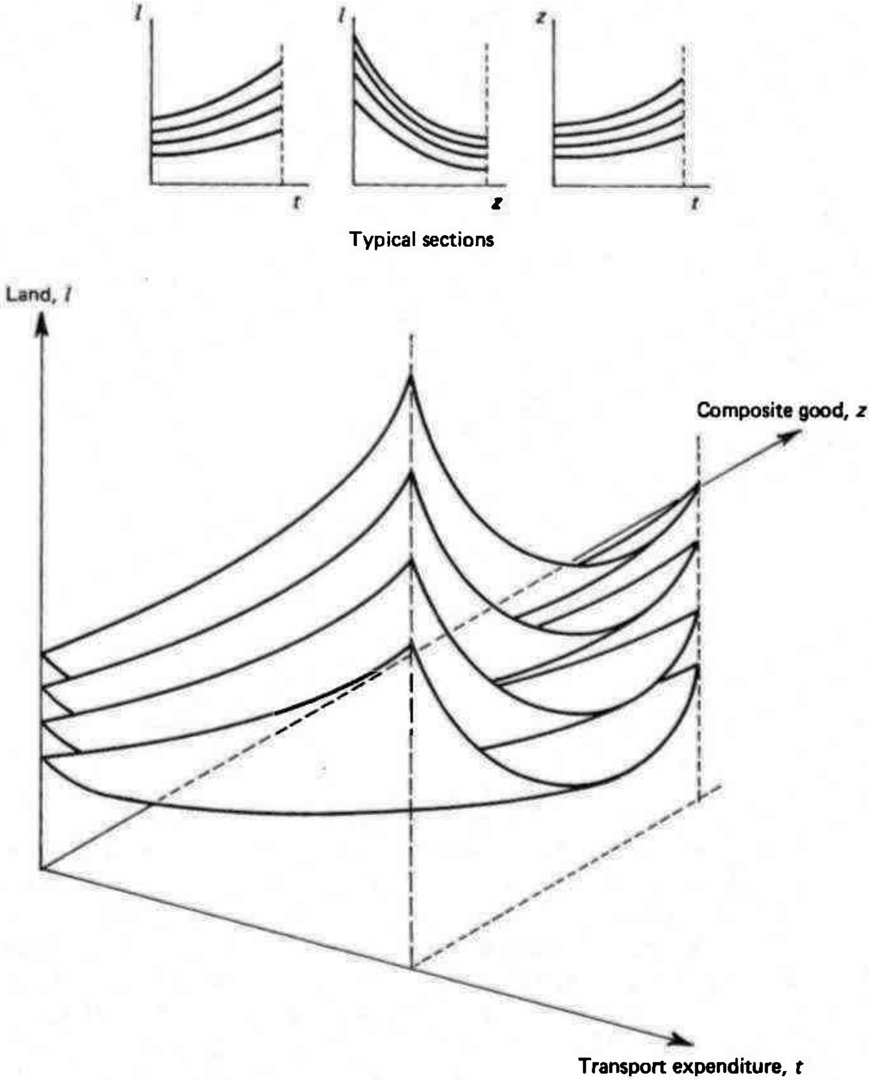


Figure 2.10. Indifference map in Alonso's model

same type will have an infinite number of parallel indifference curves. Curves higher up in the diagram will represent higher levels of utility.

The second diagram in figure 2.10 shows indifference curves between land and the composite good. In this case both have a positive utility, so that a small amount of land with a large amount of composite good will be equally preferable to a large amount of land with a small amount of z. Finally, the third diagram shows the way in which the composite good can be combined with transport to produce the same utility. Here, again, a positive utility is combined with a negative one, so that a similar shape to the first diagram results.

When the three indifference curves are combined, a bowl-shaped indifference surface results, also shown in figure 2.10. Any point in the surface will produce the same utility to individuals. Bowls closer to the origin will represent lower levels of satisfaction than those further away from it.

Analytically, utility can be expressed as a function of the consumption of land, transport and composite good:

$$u = f(l, t, z) \tag{2.17}$$

This function will be maximum when:

$$\mathrm{d}u = 0 = u_l \mathrm{d}l + u_t \mathrm{d}t + u_z \mathrm{d}z \tag{2.18}$$

where u_l, u_t and u_z represent utility with respect to land, transport and composite good respectively. According to micro-economic theory, at equilibrium the marginal rate of substitution between any two goods will be equal to the ratio of their marginal costs. In this case:

$$\frac{u_l}{u_z} = \frac{r}{p_z} \tag{2.19}$$

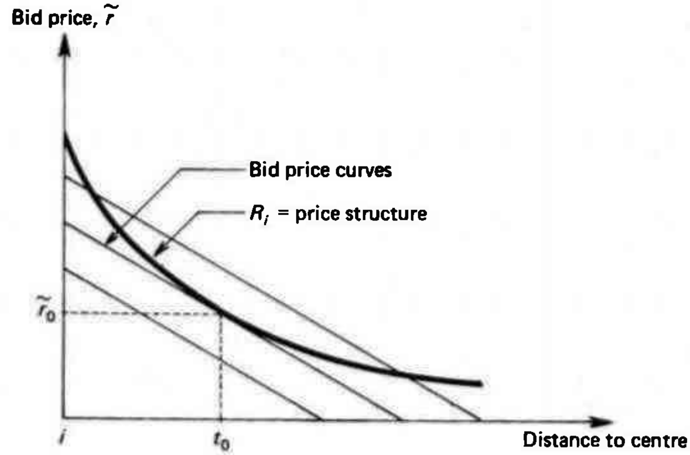
and

$$\frac{u_t}{u_z} = \frac{\left(1 \frac{\mathrm{d}r}{\mathrm{d}t} + \frac{\mathrm{d}c}{\mathrm{d}t}\right)}{p_z} \tag{2.20}$$

Both u_l/u_z and u_t/u_z represent marginal rates of substitution, and the right-hand sides of (2.19) and (2.20) represent cost ratios. In (2.19), the ratio between the utilities with respect to land and the composite good is equal to the ratio of their corresponding prices, that is, the price of land, r, and the price of the composite good, p_z. In (2.20) a similar ratio is established between transport and the composite good. l is the amount of land consumed. In this case $(l(\mathrm{d}r/\mathrm{d}t) + (\mathrm{d}c/\mathrm{d}t))$ represents the marginal cost of moving away from the centre. As the individual moves further out, the price of land decreases at a rate of $\mathrm{d}r/\mathrm{d}t$, and the cost of transport, c, increases at a rate $\mathrm{d}c/\mathrm{d}t$. If (2.19) is solved simultaneously with (2.18) and the budget constraint equation (2.16), the optimal combinations of land, transport and the composite good can be obtained.

At this point Alonso introduces the concept of *bid price curve*, describing the price an individual is willing to pay as he moves away from the city centre to obtain the same level of satisfaction. This is equivalent to a demand price at constant utility. At the centre itself, bid prices are maximum, and they will decay towards the margin. This is shown in figure 2.11: near the centre at a distance t_1 an individual is willing to pay a bid price of $\bar{r}_1$ and further away at a distance t_2 the corresponding bid price is $\bar{r}_2$. Both situations will produce the same utility for the individual and in general, all points along the bid price curve will represent the same utility. Each individual will have an infinite set of parallel bid price curves, as shown in figure 2.11; in this case, lower bid price curves imply greater satisfaction, because the same can be obtained with less money. Note that a bid price $\bar{r}$ is not necessarily equal

Figure 2.11. Bid price curves for an individual or firm in Alonso's model



to the final market equilibrium price r, since bids do not have to be successful.

At equilibrium, the price structure (land rent at every point) and the bid price curve must be tangent. This is because if they should cross it would mean that the price structure was coming in contact with some lower bid price curve, which by definition would be preferable. Equilibrium price is point $(t_0, \bar{r}_0)$ in figure 2.11, which is the point where the slope of both curves is equal, that is, where $dr/dt = d\tilde{r}/dt$.

Bid price curves of different individuals will vary in slope according to their behaviour with respect to the cost of transport and the ease with which they are willing to substitute with the other goods l and z. Figure 2.12 shows a case with two users 1 and 2. User 1 has very steep bid price curves, which shows that he is rather anxious to live near the centre, either because he is poor and cannot afford high transport costs

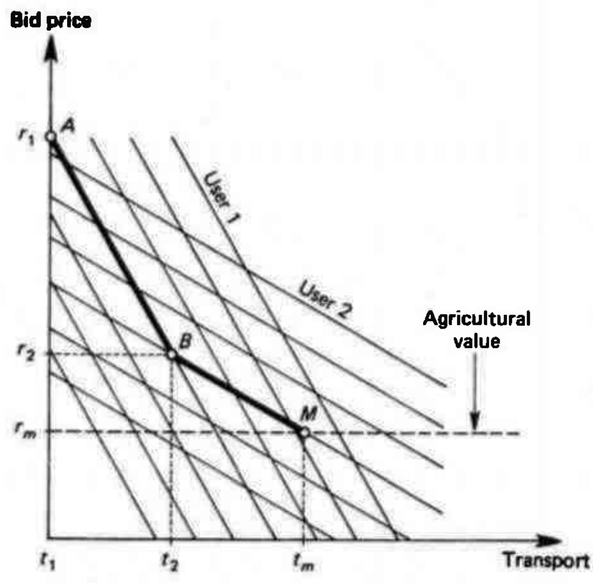


Figure 2.12. Chain of bid price curves for two users in Alonso's model

and would rather consume a small amount of expensive land, or because the individual, say, is a commercial firm that depends on a central location to attract clients. In both cases their utility decays rapidly as they move away from the city centre. In contrast, bid price curves of user 2 are not so steep, because it is, say, the case of a wealthier household that can afford higher transport costs and enjoys larger plots of land and the pleasures of suburban life. As a result, point A will be the *equilibrium price* of user 1 with a value of (r_1, t_1). User 1 will locate, then, at a distance t_1 from the centre and will use all of the land from A to B. Rent r_1 will be the equilibrium price of user 1 and r_2 will be his *price at the margin*. At the same time r_2 will be the equilibrium price of user 2, and if there are no more users, r^m is the price at the margin of user 2 and also the margin of the city, at a distance t_m.

Two conclusions can be drawn at this point:

(a) users with steeper bid price curves will locate nearer to the city centre, and
(b) the price at the margin of one user corresponds to the equilibrium price of the user immediately following him.

Up to now, the price structure of land has been assumed to be known, but the purpose of the model is, precisely, to determine the resulting land prices in every location, given that in the city there is a certain population of households and firms, each one with characteristic bid price curves. As in the case of the model of Von Thünen, the problem is non-linear, and Alonso resorts to an iterative solution. Alonso does provide an analytical solution using differential equations, but it can only be used after an approximation has been found using the iterative method.

Alonso's iterative solution starts from the most central location, where an initial guess for the price of land r_1 is adopted. The user with the steepest bid price curve will locate at the most central location, indicated in figure 2.13 as user 1 with location t_1. The next user, that is, the user with the second steepest bid price curve, will take the next available location, t_2, which is the one that comes immediately after the land requirements of the first user, l_1, have been satisfied. The budget equation (2.16) can be used to determine the radius of a circle of an area of l_1, given the initial guess of the price of land in the centre.

This gives rise to the point in the diagram with a rent value r_2 and distance to the centre t_2. The value of r_1 is the equilibrium price of user 1, and the value of r_2 is simultaneously the equilibrium price of user 2 and the price at the margin of user 1. Thus, a chain of bid price curves begins to take shape.

The algorithm continues by processing each user in turn in order of the steepness of their bid price curves. When the last user has been dealt with, say user n, a land price at his margin, r_{n+1}, can be estimated. The price of land beyond the periphery of the city, however, can be estimated independently: if agricultural firms have been included in the list of users, then $r^{m+1} = 0$, otherwise r^{m+1} = agricultural price. Should the

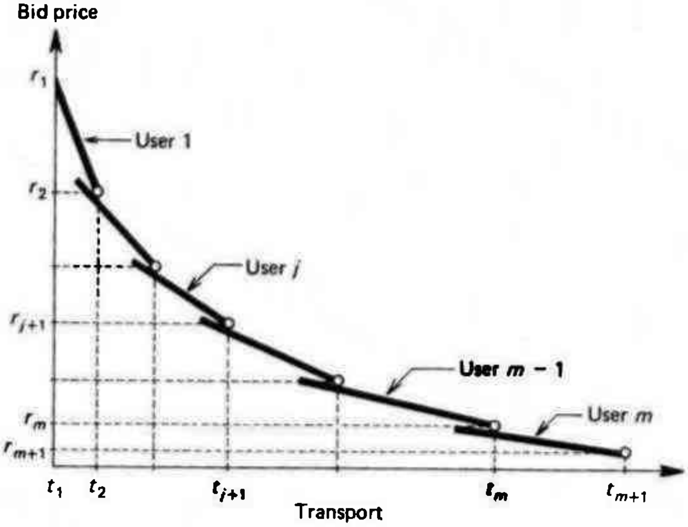


Figure 2.13. Chain of bid price curves for many users in Alonso's model

two estimates of the value of land at the edge of the urban area coincide, it would mean that the initial guess of r_1 was correct. Should there be a difference, positive or negative, the initial guess of r_1 must be re-estimated accordingly, calculating the chain all over again, checking the value at the margin against the independent estimate. This iterative process must be repeated until the correct value is obtained within a pre-defined tolerance.

2.6 Multiple market areas: the models of Christaller and Lösch

The work of Christaller and Lösch is based on Von Thünen's original model, but provides an explanation of how more complex regions are formed, with multiple market centres, and also of how these regions relate to each other. Their explanation is generally known as *central place theory*.

The explanation begins by making a number of simplifying assumptions, such as the existence of a large featureless plain, where an evenly distributed population of self-sufficient agricultural producers have settled. Each farmer produces commodities for his own consumption, such as, for example, beer. Technological development reaches a point where some farmers are able to organise beer production on a larger scale. Since the region is perfectly homogeneous, the first farmer to do so can have any randomly determined location.

Whatever the point in space, the first question a potential producer will have to face is how much beer he will be able to sell. Economies of scale will have a stimulating effect on the demand for beer, but the cost of transport from the factory to the consumers will have the opposite effect. Also, depending on the price, consumers will demand variable quantities, such that a demand function can be established, as in figure 2.14(a).

At the factory, the cost of beer would be the cost of production *OP*,

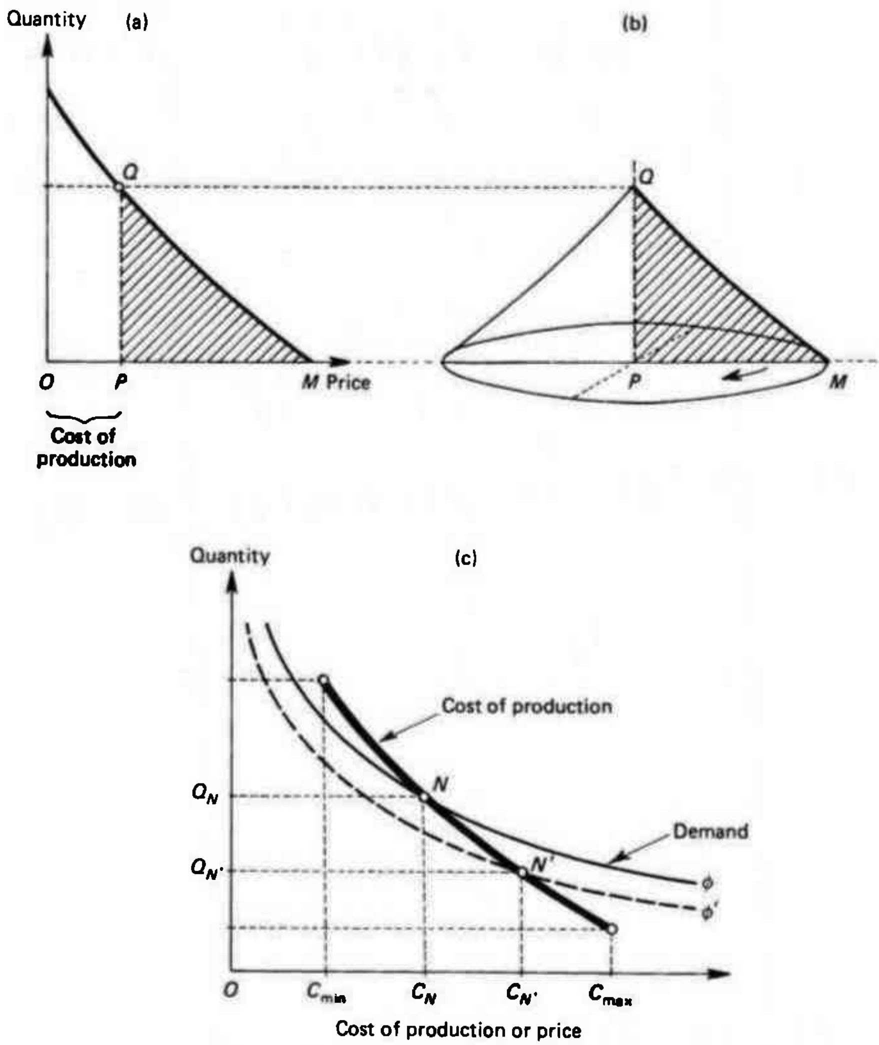


Figure 2.14. Formation of a market area in central place theory

where consumers are willing to buy the maximum amount of beer *PQ*. As the distance to the factory increases, the price also increases due to transport costs, forcing demand down, until point *M* is reached, where transport costs are such that no more beer can be sold. *PM* is then the radius of the *market area* of a production unit located in *P*. By rotating triangle *PQM* around *PQ*, a cone-shaped volume will emerge as in figure 2.14(b), which, if multiplied by the density of the population, will represent the total amount of beer sold.

In the explanation above it was assumed that *OP* was the cost of production. However, the cost of production will, in turn, depend on the size of the demand. Figure 2.14(c) shows two curves: a cost of production function and the total demand function ϕ. The cost of production function varies from a maximum C_{max} to a minimum C_{min}, depending on the amount of beer sold, and demand will decay as the price of beer rises. Our farmer will not be able to produce beer if the cost of production curve lies under the demand curve. At point N, however, both curves intersect, so that he will be able to sell an amount Q_N at a price C_N. Q_N should be the total amount of beer sold, previously determined in figures 2.14(a) and (b).

Assume, now, that a second producer of beer emerges in another (randomly selected) point in space. If this second producer is at a distance larger than twice PM, another market area of the same size will emerge, as in figure 2.15(a). Third and fourth producers can emerge next as shown in figure 2.15(b), this time overlapping with the previous producers. When two market areas overlap, demand will shrink for the two producers involved. This is shown in figure 2.14(c) above, where the demand curve ϕ will shift to ϕ', moving the equilibrium point to N'. The bigger the overlap, the greater the shift of the demand curve.

As more and more producers emerge in the landscape, the number

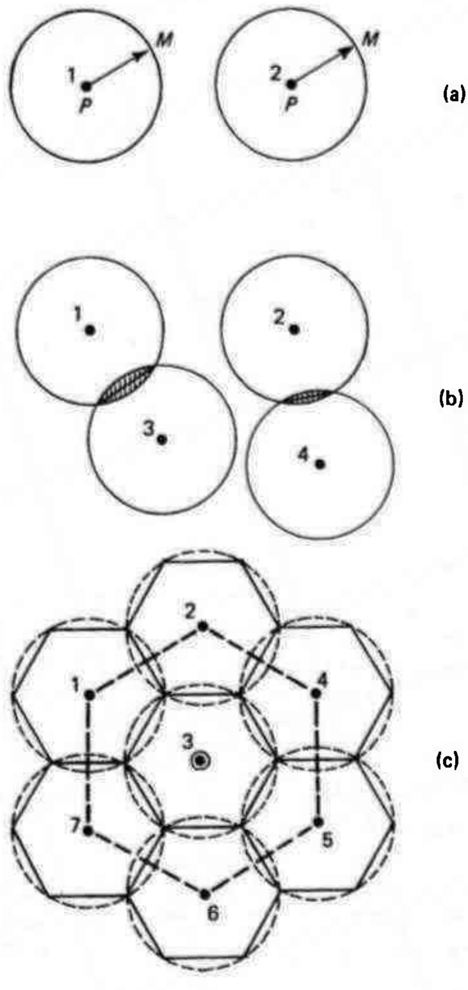


Figure 2.15. Formation of a system of market areas in central place theory

and size of the overlaps grow, until for some the demand curve will shift to such an extent that it will no longer intersect the cost of production curve. Such producers will close down. The process will continue until a state of equilibrium is reached, where all producers are located as close as possible to each other, and no more producers can leave or enter the system. In equilibrium, all cost of production curves will be tangent to the corresponding demand curves, and as Christaller demonstrated, market areas will be of a regular hexagonal shape. This is shown in figure 2.15(c). In this system, a fine mesh of small roads will be required for the transportation of beer from producers to consumers.

Next assume that, since there is now a large number of beer producers, someone decides to establish a factory of barrels. The most logical location will be immediately close to one of the beer producers, and since they are all of the same size and evenly distributed in space, the selection can be random, say 3 in figure 2.15(c). The same analysis applied to the producers of beer can be applied now to barrels, with different parameters. This will result in market areas of bigger size, as shown in figure 2.15(c), which, at the same time, must match a discrete number of beer producers. In other words, production of barrels must adjust its equilibrium point to maximise the number of coincidences with the beer producers. This leads to a higher-order hierarchy of market areas, where some points in space will now house both a beer and a barrel factory. Also of importance is the fact that some of the previous small roads will now carry beer and barrels, and consequently will grow in importance.

The region will grow in complexity as more and more commodities are introduced. The result will be a large number of regular hexagonal patterns of different hierarchical orders, laid on top of each other. Some points in space will have only one or two producers, but others will accumulate a large number, becoming important metropolitan areas. Similarly, some roads will carry only a small number of commodities short distances, becoming local roads, and others will carry a large number of commodities long distances, becoming regional motorways of national importance.

2.7 Some conclusions about micro-economic models

The models described above have many features in common. Firstly, they explain land use as the result of a market mechanism, in which individual households and firms compete for space, generating an equilibrium pattern of land rent. At the same time, equilibrium prices allow for the optimum allocation of land to households and firms, and these, in turn, maximise their utilities.

The model of Von Thünen, in spite of the fact that it was developed so many years ago, remains the most general proposition, because it equilibrates not only a land market, but a market of commodities as well. Von Thünen's basic model has been extended in this chapter to include demand elasticities of land and commodities. The result is a model which is even more general, and that contains all the variables required to give a broad and simplified explanation of how a spatial and economic system works.

The models of Wingo and Alonso update Von Thünen's proposition by incorporating many elements from modern micro-economic analysis. Conceptually, however, the only element that they add to the extended version is the budget constraint, with two elements in Wingo and three in Alonso. A budget constraint equation is, however, implicit in the extended version of Von Thünen's model through cross-elasticities, and could be easily applied to a hypothetical urban case. The reader can easily check this by mentally or practically applying program VONTH to an urban case: instead of, say, three commodities, commercial firms and two types of households could be introduced. An interesting corollary to such an application would be that a specific relationship could be established between land rent and salaries. In fact, if households have to pay more for land, they become a more expensive 'commodity'.

The contributions of Christaller and Lösch are important because they provide an explanation of the way in which multi-centre regions can be formed. In this case, each commodity gives rise to its own pattern of location, a network of market areas. In turn, each pattern is conditioned by the others, forming hierarchies of patterns and transport networks of great potential from the point of view of theoretical development. Ideally, another computer program could be written to demonstrate the formation of complex multi-level regions, using random elements to create a different landscape in every run. It must be pointed out that the random elements included in the explanation of central place theory are not part of the original work of Christaller and Lösch, and hence must be considered as another extension.

However important these theoretical developments, the achievements of spatial micro-economic models have not been matched by practical applications. As Wilson (1974) points out, with the possible exception of the Herbert and Stevens model (1960), 'no very exciting operational models have developed from the work. Many insights and qualitative analyses have been obtained, but few effective models. One possible reason for this is that most economists in the field, having established their theoretical framework, then resort to essentially linear or log-linear econometric models for their empirical work.' Alonso (1964) does, in fact, present a very simple linear model as empirical evidence of an otherwise highly elaborate theory.

Even if this restriction imposed by the tradition of econometrics is of importance, there are many other restrictions that can be mentioned, particularly the way in which space and activities are represented. Because micro-economic models treat space as a continuous variable, it is all but impossible to represent the variety and richness of the urban and regional geography. Apart from Wingo and Alonso, there are other authors that have devoted considerable effort to the representation of different shapes in their continuous formulations, only to end up with a few workable solutions difficult to find in reality, such as ovals, star shapes, squares and so on. A further difficulty is that in this approach it is impossible to model more than one market-place or centre of employment. Consider equation (2.13) in Wingo's model as an example, where rent in any point j is calculated as the savings in transportation with respect to the margin, i.e. $R_j = c_{im} - c_{ij}$. If two centres of employment are introduced, the rent and density surfaces will resemble a two-pole circus tent; the model will now have two margins, m_1 and m_2, two c_{ij}s, and four c_{im}s, making the whole model collapse.

Furthermore, it is difficult to model the individual behaviour of households, firms and landowners in any practical sense, since thousands or even millions of them can be found in any city or region. In real case studies, aggregation is inevitable. However, as Alonso (1964) himself points out, in the micro-economic approach it is impossible to aggregate individual demand curves in the same way as in classical equilibrium theory. Since demand curves for land for the same individual will vary with location, it is impossible to know which curves must be added in the aggregation process.

It is clear from the arguments above that, if they are to have any practical application, micro-economic models require fundamental modifications. This is important not only for professional practice, but for theoretical development as well. In chapter 1, models were defined as testable propositions derived from theories. It is only when theories become testable models subject to refutation, that they can be assessed and improved.

In the next chapters, different approaches and theoretical frameworks will be discussed, particularly spatial interaction. It will be seen that, by solving the problem of aggregation, many of the deficiencies mentioned above can be removed, thus opening a broad area for research and development.

3

Spatial interaction models

The micro-economic models described in the previous chapter can be defined as belonging to a disaggregated approach, because the analysis centres around the behaviour of individual units. The approach taken by spatial interaction models, by contrast, can be defined as aggregate, because both space and activities are grouped into discrete categories. Instead of analysing particular points in space, zones containing a large number of activities are defined. Activities, on the other hand, are aggregated into groups, and it is assumed that all the individual members of a group have similar characteristics.

3.1 Origins and development of spatial interaction models

There is a considerable amount of literature on spatial interaction models, so that only a brief review is presented in this section. Excellent reviews are contained in Lee (1973), Batty (1975, 1976), and Baxter (1976). In spite of earlier work by Reilly (1931), Hoyt (1939), Stewart (1948), Zipf (1949), Converse (1949), Clark (1951), Isard (1956), and others, most authors agree on placing the origins of modern spatial interaction models in 1959 with the work of Hansen. After this, a flurry of research activity took place.

There are many factors which led to the rapid growth of this area of research, but perhaps the most important one is that spatial interaction models are easy to apply to real cases, producing useful and realistic results. Another factor is that the spatial interaction approach is particularly relevant to transportation analysis. During the 1960s and early 1970s, as a result of rapid urban growth and increase in car availability, many transport-related projects were implemented, such as motorways and mass transit systems, which meant that spatial interaction models were in great demand.

The first spatial interaction models were mainly based on a gravitational analogy, something which derived naturally from the aggregate approach. Instead of looking at the individual 'molecules' of an urban area, spatial interaction was more interested in the behaviour of whole 'masses', and the relationship among them. The original gravitational analogy, as more empirical testing was carried out, gave way to a proper urban or regional theory.

One of the first steps in this direction was the work of Hansen (1959), who, still using the gravitational analogy, elaborated on the location of residents as a function of accessibility to employment. Then Huff (1962, 1963) made an important contribution by interpreting the basic gravity model in economic terms and probabilities. Lowry (1964) achieved a landmark in the history of spatial interaction models, by using economic base principles and introducing a multiplier to provide a more comprehensive explanation of the urban structure. Lowry's work was then improved by the work of Rogers (1968) and Garin (1966) on matrix methods. Spatial interaction drifted even further away from its original gravity formulation with the important work of Wilson (1967, 1970, 1974) on entropy maximisation. This method created the basis for the development and implementation of numerous operational models, such as those by Echenique (1968) and Batty (1976).

The possibility of applying spatial interaction models to real cases has been further improved by the development of calibration techniques to estimate the various parameters involved, particularly with the work of Angel and Hyman (1971), Baxter and Williams (1975). The possibility of disaggregating Lowry's original model has also been a constant preoccupation of investigators, and good examples are the work of Baxter and Williams (1973) and Wilson (1974).

3.2 Basic concepts

In spatial interaction models, land used by activities is defined as aggregate units of space or *zones*, containing a certain number of activities within them. These aggregates interact, generating flows of different kinds, which can be of a concrete nature – such as trips, migrations, movements of commodities, etc – or of a more abstract nature, such as dependencies, diffusions, opportunities, etc. Each zone is described in terms of a number of attributes. The zones are linked to each other by means of infrastructures or networks, depending on the nature of the flows.

The gravity form of a spatial interaction model states that interaction between any two zones is proportional to the number of activities in each zone (masses), and inversely proportional to the friction imposed by the particular infrastructure that connects them. The most simple formulation is then:

$$F_{ij} = gM_iM_jf(c_{ij}) \tag{3.1}$$

where F_{ij} is the magnitude of the flow between zones i and j, M_i and M_j are the number of activities in zones i and j, and $f(c_{ij})$ is a (negative) function of the friction imposed by the infrastructure connecting i to j, measured in terms of distance, time, cost, or any other variable. Term g is a constant that transforms the activity units into the flow units.

If the system is composed of more than two zones, flows between any particular pair of zones must be restricted by the combined effect of all other zones present in the system:

$$F_{ij} = g\frac{M_iM_jf(c_{ij})}{\Sigma_jM_jf(c_{ij})} \tag{3.2}$$

The denominator represents the effect of all zones in the system, including i and j. The model can be used in several ways. If the main interest are the flows themselves, as in a transport demand model, F_{ij} would represent trips, M would represent attraction and production variables, and g a trip generation factor. If, however, the model was intended for the simulation of the location of activities, F_{ij} would become an abstract flow, representing activities that are generated in i and located in j; once flows have been generated for all pairs in the system, they can be added with respect to the origin to give the total amount of activity in each zone, X_j:

$$X_j = \Sigma_i F_{ij} \tag{3.3}$$

Alternatively, the model can be used to measure the 'potential' of a particular zone with respect to all others. This kind of analysis can be useful for diagnosis, to determine the development potential of a zone, or for design purposes, to determine the optimal location of an industry with respect to a particular market. The potential of a zone, V_i, can be calculated as:

$$V_i = g\Sigma_j M_j f(c_{ij}) \tag{3.4}$$

A potential model may also become an activity location model:

$$X_j = X_* \frac{V_j}{\Sigma_j V_j} \tag{3.5}$$

where X_* is the total number of activities to allocate, and X_j is the number of activities located in j.

3.3 The entropy maximising derivation

The formulations discussed above are based on rather simple, heuristic assumptions. Wilson (1970) introduced entropy maximising methods, which gave spatial interaction models a new and more general theoretical framework. The concept of entropy was originally developed in statistical mechanics and later proposed as a general theory of information applicable to most systems. There are many ways of interpreting entropy: it can be related to probability and uncertainty, to a probability distribution or to Bayesian statistics. It must be pointed out that the introduction of entropy to urban and regional theory does not mean that yet another analogy is being adopted, since entropy maximising is a very general method. As March (1972) points out, 'to say that if the maximum entropy approach is applied to urban studies it is in effect treating people as though they were gas molecules is exactly as meaningful as to say that if the approach is applied to the study of gases it is in effect treating gas molecules as though they were people'.

Consider a simple example to illustrate the main components of this method. Imagine a completely empty bottle which is suddenly filled with a certain amount of gas from another smaller bottle. On entering the bottle, the molecules of gas will move in all directions, filling the bottle completely. The fact that the gas completely fills the space within the bottle means that the molecules have spread as evenly as possible, arriving at the most random distribution, where all possible locations are equally likely.

Based on the above example, a few definitions can be introduced. Firstly, the molecules and the container form a system which moves from an *initial state* to a *final state*. The system is in its initial state when the molecules are in the smaller bottle, and arrives at the final state once the molecules have entered and are stable in the bigger bottle. The problem is to explain the path from the initial state to the final state, given that we have complete knowledge of where the molecules were in the small bottle but ignore their final location in the bigger bottle.

Imagine dividing each bottle into two compartments. Furthermore, assume that the total number of molecules is six, and that the initial location of molecules in the small bottle is four in compartment 1 and two in compartment 2. As they enter the bigger bottle, the molecules can move freely to any compartment. The problem is to find the most probable distribution of molecules at the end of the process. The following table shows all possible outcomes.

	Possible final states						
	1	2	3	4	5	6	7
Compartment 1	6	5	4	3	2	1	0
Compartment 2	0	6	2	3	4	5	6
Number of permutations	1	6	15	20	15	6	1 (total 64)

In order to produce final state 6–0, only one possible move could have happened: the molecules of compartment 2 in the initial state moved to compartment 1. Final state 5–1, however, can be obtained in several ways, such as one of the molecules in 2 moves to 1, or both move to 1 and one of the molecules of 1 moves to 2, and so on. In total, there are six possible ways of generating final state 5–1, and similarly fifteen to produce 4–2, and so on. The total number of permutations is $2^6 = 64$. In answer to which is the most probable final state, it is clear that it is 3–3, because it is obtained with the largest number of permutations, twenty times more likely than the least likely.

From this point of view, entropy is the degree of likelihood of the final state of a system. Outcomes can be called *macro-states* of the system, and each one of the permutations (64 in the example) can be called *micro-states*. Each macro-state, then, can be obtained with one or more micro-states, and the most probable macro-state is the one which is associated with the maximum number of micro-states. However, as the probability of a macro-state increases, the uncertainty as to which micro-state gave rise to it also increases. In the example, there is no uncertainty as to which micro-state led to 6–0 because there is only one possibility, but when 5–1 is considered there are six possible ways of achieving it, and there are no grounds for choosing among them. In the extreme, entropy is maximum when uncertainty is also maximum.

In the example, the degree of 'dominance' of the most likely result over the least likely was 20 to 1. When more complex systems are considered, that is, when there are more molecules and more compartments, the degree of dominance can increase considerably. It is very often possible to introduce to the system additional knowledge in the form of constraints. In the example, the fact that the initial state was 4–2 constrains the number of permutations. If we knew, for instance, that compartments can contain only up to five molecules, then macro-states 6–0 and 0–6 would be ruled out, reducing the number of permutations to 62, thus increasing the probability of macro-state 3–3. As a conclusion, as the number of elements and the number of restrictions increase in a system, the dominance of the most likely outcome also increases, making it easier to predict than a simple system.

In order to apply these concepts to spatial interaction, Wilson (1970) considers the case of a home-to-work origin–destination table. A particular distribution of people travelling from home to work can be described as a matrix T_{ij}. Assume that the number of workers living in each zone, O_i, is known, as well as their places of work, D_j. There are many possible trip matrices that add up to the number of origins and destinations. The following are two examples:

$$\begin{array}{ccccc} & & D_j & & D_j \\ & \begin{array}{|ccc|}\hline 3 & 1 & 1 \\ 5 & 2 & 1 \\ 2 & 2 & 4 \\ \hline \end{array} & \begin{array}{|c|}\hline 5 \\ 8 \\ 8 \\ \hline \end{array} & \begin{array}{|ccc|}\hline 4 & 0 & 1 \\ 3 & 2 & 3 \\ 3 & 3 & 2 \\ \hline \end{array} & \begin{array}{|c|}\hline 5 \\ 8 \\ 8 \\ \hline \end{array} \\ O_i = & \boxed{10 \quad 5 \quad 6} & \boxed{21} & \boxed{10 \quad 5 \quad 6} & \boxed{21} \end{array}$$

In both cases, the number of origins and destinations is the same, a total of 21. These tables are only two of a large number of possible T_{ij} matrices that add up to the given origins and destinations, and, in turn, there are many ways to combine homes and works to produce a single matrix. Each T_{ij} matrix represents a macro-state that can be produced with a large number of micro-states. The problem is to find the T_{ij} that is associated with the maximum number of micro-states, assuming a priori that they are all equally likely, and that the following can be introduced as constraints:

$$\Sigma_j T_{ij} = O_i \tag{3.6}$$

$$\Sigma_i T_{ij} = D_j \tag{3.7}$$

and

$$\Sigma_i \Sigma_j T_{ij} c_{ij} = C \tag{3.8}$$

Constraints (3.6) and (3.7) were part of the original definition of the problem, i.e. the sum of the columns of matrix T_{ij} must equal the given origins (3.6) and the sum of the rows must equal the given destinations (3.7). A new constraint (3.8) has been added, which states that there is a total cost of transport in the system, C, and that this should equal the sum of all trips multiplied by the cost of transport for each origin–destination pair c_{ij}. This is equivalent to the introduction of a total energy constraint in the previous example: the molecules are free to move from one compartment to another, but in doing so they spend energy, the sum of which is fixed. This total amount of money being spent on transport gives rise to a higher level macro-state of the system, because many trip matrices can satisfy this constraint.

The complete derivation of the entropy maximising model can be found in Wilson (1970, 1974), and here only a short description is given. The method must first determine an expression to calculate the number of combinations that can give rise to a trip matrix, and then maximise it subject to the constraints.

Defining T as the total number of trips ($T = \Sigma_{ij} T_{ij}$), the total number of combinations that can give rise to T_{ij} can be calculated as follows: first select T_{11} from T; then T_{12} from $T - T_{11}$; and so on. In other words, the number of possible assignments is the number of ways of selecting T_{11} from T, which we will denote as ${}^{T}C_{T_{11}}$, times the number of ways of selecting T_{12} from $T - T_{11}$, denoted as ${}^{T-T_{11}}C_{T_{12}}$, and so on. If we define W as the total number of possible assignments, then:

$$W = {}^{T}C_{T_{11}} \quad {}^{T-T_{11}}C_{T_{12}} \quad {}^{T-T_{11}-T_{12}}C_{T_{13}} \ldots \tag{3.9}$$

so that

$$W = \frac{T!}{T_{11}!(T - T_{11})!} \frac{(T - T_{11})!}{T_{12}!(T - T_{11} - T_{12})!} \cdots = \frac{T!}{\Pi_{ij} T_{ij}!} \tag{3.10}$$

The T_{ij} which maximises W subject to constraints (3.6), (3.7) and (3.8) can be obtained with the standard method of constrained maximisation. The Lagrangian $\mathscr{L}$ has to be maximised:

$$\mathscr{L} = \ln W + \Sigma_i {}^1\tau_i(O_i - \Sigma_j T_{ij}) + \Sigma_j {}^2\tau_j(D_j - \Sigma_i T_{ij})$$
$$+ \beta(C - \Sigma_{ij} T_{ij} c_{ij}) \tag{3.11}$$

where ${}^1\tau_i$ is the Lagrangian multiplier associated with constraint (3.6), ${}^2\tau_j$ is the multiplier associated with (3.7), and β with (3.8). The T_{ij}s which maximise $\mathscr{L}$ and which are therefore the most probable distribution of trips are the solutions of:

$$\delta\mathscr{L}/\delta T_{ij} = 0 \tag{3.12}$$

together with constraint equations (3.6), (3.7) and (3.8). Maximising ln W rather than W, and using Stirling's approximation:

$$\delta\mathscr{L}/\delta T_{ij} \quad = \quad -\ln T_{ij} - {}^{1}\tau_i - {}^{2}\tau_j - \beta c_{ij} \tag{3.13}$$

Equation (3.13) will be equal to zero when:

$$T_{ij} \quad = \quad \exp(-{}^{1}\tau_i - {}^{2}\tau_j - \beta c_{ij}) \tag{3.14}$$

Substituting (3.14) in the constraint equations (3.6) and (3.7) and re-arranging:

$$\exp(-{}^{1}\tau_i) \quad = \quad O_i[\Sigma_j \exp(-{}^{2}\tau_j - \beta c_{ij})]^{-1} \tag{3.15}$$

and

$$\exp(-{}^{2}\tau_j) \quad = \quad D_j[\Sigma_i \exp(-{}^{1}\tau_i - \beta c_{ij})]^{-1} \tag{3.16}$$

For convenience define:

$$A_i \quad = \quad \exp(-{}^{1}\tau_i)/O_i \tag{3.17}$$

and

$$B_j \quad = \quad \exp(-{}^{2}\tau_j)/D_j \tag{3.18}$$

Then:

$$T_{ij} \quad = \quad O_i D_j \exp(-\beta c_{ij}) A_i B_j \tag{3.19}$$

where

$$A_i \quad = \quad [\Sigma_j B_j D_j \exp(-\beta c_{ij})]^{-1} \tag{3.20}$$

and

$$B_j \quad = \quad \Sigma_i A_i O_i \exp(-\beta c_{ij})]^{-1} \tag{3.21}$$

In this way A_i satisfies constraint (3.6) and B_j satisfies (3.7). Equations (3.19), (3.20) and (3.21) represent the entropy maximising spatial interaction model in its most general form. Wilson, however, derived four different cases, depending on the available information:

(a) Doubly constrained. This corresponds to the case of maximum information, because both the origins O_i and destinations D_j are known. This is the case described above in equations (3.19), (3.20) and (3.21), subject to constraints (3.6) and (3.7). Note that terms A_i and B_j are mutually dependent, so that they have to be solved iteratively: first make all B_j equal to 1 in (3.20) and calculate the values of A_i; replace these in (3.21) to obtain new values of B_j; repeat this process until numerical equilibrium is reached.

(b) Origin constrained. In this case only the origins of the flows are known, but not the destinations. Only restriction (3.6) then holds, and term D_j must be replaced by a hypothesis indicating the attractiveness of the destination zone, W_j. Then:

$$T_{ij} \quad = \quad O_i W_j \exp(-\beta c_{ij}) A_i \tag{3.22}$$

where

$$A_i = [\Sigma_j W_j \exp(-\beta c_{ij})]^{-1} \tag{3.23}$$

(c) Destination constrained. Only restriction (3.7) holds:

$$T_{ij} = W_i D_j \exp(-\beta c_{ij}) B_j \tag{3.24}$$

where

$$B_j = [\Sigma_i W_i \exp(-\beta c_{ij})]^{-1} \tag{3.25}$$

(d) Unconstrained. This corresponds to the situation of minimum information where neither origins or destinations are known, and hence no constraints are applicable. Then:

$$T_{ij} = W_i W_j \exp(-\beta c_{ij}) \tag{3.26}$$

3.4 Some specific forms of spatial interaction models

Applications of spatial interaction models are numerous, so that only a few examples are presented here. First those models related to the simulation of flows are presented, followed by models related to the location of activities. Formal representation of these models has been changed from the original to adapt them to the notation of the preceding sections.

3.4.1 Models of flows

One of the best-known models of flows is Wilson's (1970) trip distribution and modal split model, which is based on the type (a) doubly-constrained spatial interaction model described above. The purpose of the model is to simulate the number of trips between origin zones i to destination zones j by transport mode k and population type n, T_{ij}^{kn}. Population type represents, generally, car-ownership ($n = 1$ car-owners, $n = 2$ non-car-owners). Inputs to the model are the number of trip origins by population type, O_i^n, and the number of destinations in each zone, D_j, as well as travel cost matrices by mode, c_{ij}^k. Then:

$$T_{ij}^{kn} = O_i^n D_j \exp(-\beta c_{ij}^k) A_i^n B_j \tag{3.27}$$

where

$$A_i^n = [\Sigma_j \Sigma_{k \in n} B_j D_j \exp(-\beta c_{ij}^k)]^{-1} \tag{3.28}$$

and

$$B_j = [\Sigma_i \Sigma_n \Sigma_{k \in n} A_i^n O_i^n \exp(-\beta c_{ij}^k)]^{-1} \tag{3.29}$$

The expression $k \in n$ indicates that the summation is made over all modes k available to population type n. If, say, two modes are being considered – public and private – non-car-owners are forced to choose only from the former, while car-owners are allowed to choose from both.

3.4.2 Models of the location of activities

One of the first location models was due to Hansen (1959) to simulate the location of residents in an urban area. The model assumes that there has been an increment in population, dR, which must be allocated to zones as a function of accessibility. Accessibility is measured as:

$$V_i = \Sigma_j M_j \exp(-\beta c_{ij}) \tag{3.30}$$

where M_j represents the number of attracting activities in zone j, such as the number of jobs, services, etc. To account for the size of zones, Hansen proposed the following allocation equation:

$$\mathrm{d}R_i = \mathrm{d}R \frac{L_i V_i}{\Sigma_i L_i V_i} \tag{3.31}$$

where dR_i is the increase in population in zone i and L_i is the vacant land available in zone i.

Later, Lakshmanan and Hansen (1965) proposed a model to simulate the potential location of retail services. Expressed in terms of an origin-constrained model, the flow of retail expenditure from a residential zone i to a shopping centre in j, S_{ij}, is:

$$S_{ij} = R_i e_i W_j A_i \exp(-\beta c_{ij}) \tag{3.32}$$

where

$$A_i = [\Sigma_j W_j \exp(-\beta c_{ij})]^{-1} \tag{3.33}$$

and where e_i is the average per capita expenditure of the population in zone i, and W_j is a variable representing the relative attractiveness of shopping facilities in j, generally in terms of floorspace area.

Perhaps the most popular of all spatial interaction models is that of Lowry (1964). It defines the urban system as composed of a basic employment sector, a service employment sector and a residential sector. Basic employment in each zone, ${}^bE_i$, is exogenous, and the purpose of the model is to estimate the location of residents, R_i, and of service jobs, ${}^sE_i$, that is derived from the location of basic jobs. Other exogenous variables are the land available in each zone, L_i, and an accessibility matrix or matrix of transport costs, c_{ij}.

The main contribution of Lowry is that several sub-models are linked to each other within an iterative system, thus allowing for a more complex structure. The way in which these sub-models are linked is shown in figure 3.1. Expressed in terms of a series of origin-constrained spatial interaction models, the following steps describe the necessary calculations:

(a) Add to the basic employment in each zone the service employment allocated in the previous iteration. In the first iteration, service employment to be added will be zero. If ${}^*E_i$ denotes total employment, then:

$$ {}^*E_i = {}^bE_i + {}^sE_i \tag{3.34}$$

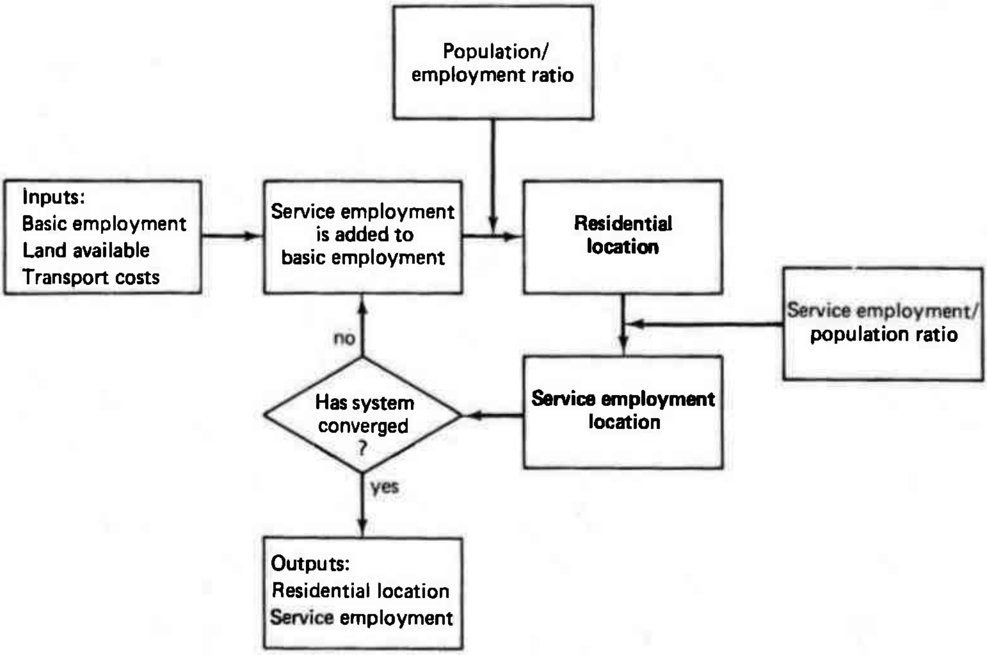


Figure 3.1. Basic structure of the Lowry model

(b) Allocate residents to zones j from work-places in i:

$$R_{ij} = {}^{*}E_i u A_i L_j \exp(-\beta^r c_{ij}) \qquad (3.35)$$

where

$$A_i = [\Sigma_j L_j \exp(-\beta^r c_{ij})]^{-1} \qquad (3.36)$$

and where R_{ij} is the number of residents in j that work in i; u is a population-to-employment ratio. Term A_i ensures that the correct number of residents is allocated, i.e. $\Sigma_j R_{ij} = E_i u$. Parameter β^r regulates the effect of transport costs on the distribution of residents; a very high value of β^r will result in the population being allocated very close to their places of work. In the limit, when $\beta^r \to \infty$, all residents will live and work in the same zone. Lower values of β^r result in the population spreading more evenly around their places of work. In the limit, when $\beta^r = 0$, they will locate entirely in proportion to the land available with equal density. Term R_{ij} represents work–home flows, and can be transformed into the total number of residents living in zone j by summing over all work-places:

$$R_j = \Sigma_i R_{ij} \qquad (3.37)$$

(c) Allocate service employment to zones j from places of residence in zones i:

$${}^{s}E_{ij} = R_i v A_i (W_j)^{\alpha} \exp(-\beta^s c_{ij}) \qquad (3.38)$$

where

$$A_i = [\Sigma_j (W_j)^\alpha \exp(-\beta^s c_{ij})]^{-1} \quad (3.39)$$

and where v represents a service-to-population ratio, W_j represents an attractor term for service employment, in terms of the service employment located in the previous iteration to represent the tendency of this activity to cluster; the value of α regulates this, and hence is sometimes called *economies of scale* parameter. If home–service flows are added with respect to the origin, the total number of service employees in each zone can be obtained:

$${}^sE_j = \Sigma_i {}^sE_{ij}. \quad (3.40)$$

Calculation goes back to step (a), where the service employment obtained in (3.40) is added to the exogenous basic employment in (3.34). In each iteration a certain number of residents and service employment is added; however, this number gets smaller and smaller, allowing for convergency after a reasonable number of iterations.

Program LOWRY can be used to experiment with this model.

A number of improvements have been made to the original Lowry model. Garin (1966) produced an interesting matrix formulation, providing useful insights and allowing for a direct solution to the iterative process by matrix inversion (this, however, is less efficient in computational terms). Echenique (1968) introduced a floorspace location model, in an attempt to incorporate a supply-side element. He observed that floorspace density tends to decay exponentially from the city centre; residential density, instead, tends to be very low near the centre because of competition with other activities, mainly service employment. Thus, Echenique splits the structure of the Lowry model into two parts: floorspace location and activity location. In this way, it is assumed that employment will determine the location of floorspace, and once a certain amount has been allocated in each zone, residential and service activities compete for its use, with the latter having preference over the former. As a result, central areas can attract a large number of residents in the first itertion, only to be pushed away from it by service employment in subsequent iterations. Program ECHE described in the appendix contains a version of this model.

There has been a number of attempts to disaggregate some of the variables of the Lowry model, introducing several income groups for residents and several types of service employment. Baxter and Williams (1973, 1975) worked on the residential model to considerable extent. They began by simply adding another index to R_{ij} in (3.35) and introducing F_j, residential floorspace, instead of L_j, land available. This assumes that all housing stock is equally attractive and available to residents of any type. The authors later dropped this assumption by introducing several types of stock in a doubly-constrained formulation.

Wilson (1974) also made an attempt to disaggregate the residential model, assuming that both the number of workers, E_j, and the number of houses, H_i, in each zone are known. A doubly-constrained model can be built of the form:

$$R_{ij} = H_i E_j \exp(-\beta c_{ij}) A_i B_j \tag{3.41}$$

where A_i and B_j are defined in the usual way. Wilson then proposes a hybrid model for urban areas in a rapid growth situation, considering four possible cases:

Case 1: a group of established residents whose job and home locations are known. This case can be solved with a doubly-constrained model of the form (3.41).

Case 2: residents whose residential location is known, but are looking for a job, dealt with an origin-constrained model.

Case 3: residents with a definite job location, looking for a house, dealt with a destination-constrained model.

Case 4: completely unsettled population looking for both jobs and houses, dealt with an unconstrained model.

From this description, Wilson derives a large number of restrictions which are then introduced into the maximum entropy derivation. The author then proposes a model that assigns workers of different income w working in zones i, to houses types k in zones j, R_{ij}^{kw}, introducing a built-in mechanism to ensure that the expenditure in housing and transport does not exceed a maximum for each income group. The price of houses is given exogenously.

The service employment model can be disaggregated also, by considering different types of services, like commercial, financial, health, education, and so on. It is also possible to distinguish between service employment generated by residents and by basic employment (the original Lowry model does this). The problem with these models is that it is difficult to calculate the many service-to-population and service-to-employment coefficients. Such information will never exist, since it is impossible to ask, say, a shopkeeper which income group does he sell to, or in what proportion. Indirect estimates, however, are possible: a single average can be calculated and then weighted by the average income of each service generating group.

3.5 Micro-economic and spatial interaction models compared

As a conclusion to this chapter, it is useful to summarise the main differences between the micro-economic and the spatial interaction approaches. These will be divided into theoretical and mathematical/operative aspects.

3.5.1 *Theoretical aspects*

Spatial micro-economic models, as has been pointed out, centre their analysis around individuals, classified as either consumers (households or firms) or suppliers (landowners or employers). These actors confront each other in the market, competing for land. After equilibrium, all land is assigned, consumers maximise their utilities subject to budget constraints, and suppliers maximise their benefits or revenues. Furthermore, land is assigned in the most efficient possible way system-wide. In other words, micro-economic models of land use take full advantage of consumer analysis and thus enjoy a sound theoretical basis.

Spatial interaction models are much more loosely structured from a theoretical point of view. In fact, they began their operational life based on very crude hypotheses and have been acquiring theory along the way, a process that continues to the present day.

These two alternative views imply a much broader and fundamental discussion. Micro-economic models start by defining highly abstract constructions, and then proceed in a deductive way in order to get closer to reality. By contrast, spatial interaction models have been essentially empiric, starting from some primitive hypotheses (like a simple analogy), testing the simulated results against real cases. From the results of the tests, new models emerge and new hypotheses replace some of the old ones, thus creating a theory in a gradual way.

The result of this is that micro-economic models have had a great influence in the development of urban theory, but have provided only few useful tools for the analysis of real cases. As Anas (1982) puts it, 'because of this, the applied fields of transportation planning and urban modelling – which stand to benefit from urban economics – have remained largely apart from it'. In other words, it can be argued that the micro-economic approach, following a traditional deductive method, can be perfectly consistent while travelling in the frictionless conditions of abstraction, but will face severe difficulties when applied to real situations.

A more specific issue in the comparison is the problem of equilibrium. It was said that micro-economic models are basically designed to deal with market systems, where through the process of perfect competition, all actors achieve their objectives, a process that is stable as long as the exogenous conditions are maintained. No such conception is found in spatial interaction models, where mathematical equilibrium is sometimes sought only in order to perform certain calculations (as in the Lowry model). In spatial interaction models there is no guarantee that, say, supply of floorspace is equal to demand, or that demand for services is adequately met by existing facilities. Nor is there any statement about prices of any of the commodities, not even of land. In fact, strong imbalances between most demand–supply elements can perfectly well take place, a condition that would be impossible in the models of Von Thünen, Wingo or Alonso.

This is a point that requires further discussion, but for the time being it must be stressed that it is both a benefit and a deficiency of spatial interaction models. It is a benefit, because real cases will be adequately simulated, whatever the degree of perfection of the market mechanisms that prevail. It is a deficiency, because, however imperfect, many market mechanisms still exist and can be of great importance, so that they should be made explicit in the models. If spatial interaction models are to be used as tools for the simulation and evaluation of the probable effects of alternative courses of action, it is quite obvious that not only a good activity location model is required: without a certain knowledge about how these locations affect the land and building markets, it is very difficult to assess the economic impact of alternative plans over the urban community.

The way in which micro-economic theory treats the behaviour of individuals has been criticised also. The assumption that all individuals are perfectly rational, in the sense that they will always choose the options that maximise their utilities, and that all individuals possess the same perfect information about the real system, has often been pointed out as unrealistic. Individuals, in fact, have only limited information, take many decisions on the basis of future, predicted events, and they do not always behave in a perfectly rational manner, among other things, because it is difficult to establish what is 'rational' in the first place, and for all individuals in identical terms.

Spatial interaction models make no assumptions about the degree of rationality of the actors involved, particularly if we consider the entropy maximising approach. In essence, it is assumed that individuals will behave in the most random possible way. By definition, the final state of the system is precisely the one in which we know the least about the behaviour of the individuals that lead to it. This would be totally unrealistic if restriction (3.8) had not been imposed on the model. What this restriction is saying is that individuals do behave in the most random possible way, but that they are only prepared to spend a predetermined total amount of money on transport. In other words, these are random molecules which, at the same time, watch their purses.

3.5.2 Mathematical and operative aspects

Micro-economic models and spatial interaction models also fall into different mathematical traditions. Micro-economic models are related to continuous formulations, such as differential equations, while spatial interaction models relate to discrete formulations and algorithms. It is argued here that the discrete approach is more flexible and powerful, and provides the key to the empirical success of spatial interaction models. The advantages of discrete formuations not only result in improved calculations, but also have an influence on the theories themselves. The fact that theories can become operational allows for multiple feedbacks between theory and reality, a point that was highlighted in chapter 1.

This discussion is of particular relevance in systems analysis where the distinction is made in terms of continuous versus discrete models. Barto (1978) argues that the tradition of classical physical science, related to continua and rates of change, and thus to differential equations, has developed in relative isolation to those scientists whose intuition has developed more directly under the influence of digitial computing, and therefore, 'find it very natural to think in terms of algorithms, data structures and automata'. Numerical analysis is the branch of mathematics that has dealt with the interpretation of traditional problems of physics, engineering, statistics, or economics, in terms of numerical calculations or algorithms that a computer can interpret and solve. In this sense, Barto argues that 'although the subject of numerical analysis focuses on the relationship between continuous and discrete methods, the perspective it provides lies thoroughly within the continuous tradition'. Numerical analysis acts as a kind of translation from continuous to discrete formulations. Take as an example a model developed by engineers to calculate the strength of the structure of a bridge, considering the characteristics of the materials, predicted loads, temperature, wind conditions, and so on. This model will be represented in continuous formulations with differential equations. Numerical analysis will produce a discrete interpretation of the continuous model, so that it can be introduced into a computer, because calculating machines are discrete by nature.

Because numerical analysis has largely been concerned with constructing discrete interpretations of continuous models, the main performance indicators relate to how close the approximation is to the original model, rather than to the behavioural properties of the real system. Questions like error bounds and the degree of approximation to the results from the original model as the step-size or mesh converges to zero, become particularly relevant. Instead, a direct interpretation of the real system to be modelled should be attempted, bypassing the continuous formulation altogether. This emerging *direct* approach places emphasis on the relationship between the model and the perceived real system, rather than on how well the original continuous formulation has been interpreted. Van Valkenburg (1974) argues that the interpretation of differential equations should be avoided: 'given the widespread availability of computers, indeed equations should seldom be used, since principles should be stated directly in algorithmic form'. When the discrete approach is taken from the beginning, the main performance indicators will be more related to the degree of disaggregation of available data or to the degree of detailing required by the analyst or by the problem itself. Even if the direct formulation of a discrete model resembles in the end the structure of a similar model to approximate a continuous formulation, the fact that it was so derived, according to Barto, 'makes a major portion of numerical analysis irrelevant'. The whole complex question of whether or not the results of a differential equation have been properly approximated does not need to be faced; rather, the crucial question concerns the validity of the model in accounting for observed data.

In many other disciplines, discrete models are being put forward for the representation of systems that traditionally have been modelled through differential equations, making them not only easier to implement by computers, but also preserving the theoretical attractiveness of the classical models. Furthermore, this recent trend not only makes it possible to operationalise classical theoretical formulations; computational models permit the exploration of entirely new classes of models. Systems analysis should permit the properties of a particular problem, rather than a priori methodological biases, to determine the analytical approach to be used; further, it should prevent such biases from determining which problems are considered for analysis.

In sum, the emerging concept of modelling and model validity holds that these can only be discussed with respect to a particular set of attributes (data-base) that are of interest to the analyst (problem), which constitutes his experimental framework. In this sense, a model can be said to be valid with respect to some experimental frameworks, but it may not be for others. Also, its discrete character does not mean that the system has the same properties in reality. For instance, a discrete time model does not imply that time itself is discrete, in the same way that a continuous time model does not imply that time is continuous. What is being suggested is that the two approaches are intellectual representations of the nature of time, and that the analyst should choose the one that better serves his interests or the one that appropriately describes the set of real attributes.

A discrete model, on the other hand, is more flexible than a continuous model. Figure 3.2 shows two forms of representing a real system. In figure 3.2(a), curve *A* represents the attributes of a real system, say a frequency distribution of a particular observed variable. Curve *B* represents a continuous representation, which attempts to replicate the real distribution as closely as possible. Curve *C* represents a discrete model that has been produced to approximate the continuous one as closely as possible. Because the choice of continuous models is restricted to a known family of distributions that can be dealt with, many of the peculiarities of the real distribution are lost. The result is that curve *B* is not a very good representation of *A*. If curve *C* has been constructed to approximate the continuous model, the result is that curve *C* must be worse than *B* in representing *A*. Furthermore, if the step

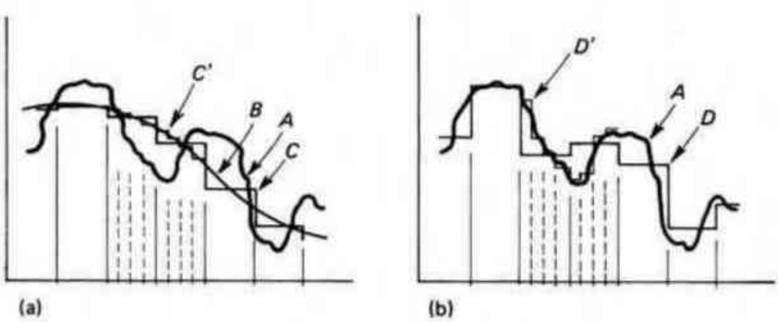


Figure 3.2. Discrete interpretation of real systems: (a) numerical analysis; (b) discrete approach. A = real attributes; B = continuous representation; C = discrete interpretation of B – thick mesh; C' = discrete interpretation of B – fine mesh; D = discrete interpretation of A – thick mesh; D' = discrete interpretation of A – fine mesh

size of the discrete interpretation is reduced, such as in curve C', no improvements can be obtained.

Figure 3.2(b) shows the advantages of a direct discrete model, bypassing the continuous model, represented by curve D. Because the discrete model is not restricted to a well-known family of continuous functions, it can achieve a much better fit to the real attributes. Flexibility is achieved because an algorithm can vary in nature as it goes along, since it is, basically, a set of calculating instructions. It can describe the real system in terms such as: first assume that the curve is exponential until a value of x is reached; then turn linear, until the value of y is reached; then jump up three times the current value and thereafter decay with a power function. Note that continuous functions can be included as part of the algorithm, further improving its capabilities. It is also quite clear that, as the step-size is reduced, such as in curve D', the model gets even closer to the real attributes. The algorithm can also vary its step-size, and include instructions such as 'if–then', 'else', 'go–to' and other logical conditions.

This discussion is particularly relevant to social sciences in general, and to urban and regional modelling in particular. The different models that have been presented in the previous sections fall into one or the other of these two traditions. In general, models based on micro-economic theory can be said to follow the continuous tradition, particularly the way in which they treat space, income, and so on (being essentially static models, though, time is implicitly treated as a discrete variable). On the other hand, spatial interaction models consider space and most of the other variables as discrete, providing an easy access to empirical testing.

This does not mean, however, that micro-economic models of land use cannot be tested because they are formulated with continuous functions. In fact, approximations can be constructed and tested, but they are bound to produce poor results. To avoid this, in the next chapter the theory of discrete choice is reviewed and taken as a basis for the construction of mixed micro-economic/spatial interaction models, in an attempt to combine the theoretically sound basis of the former with the empirical facilities of the latter.

4

Random utility theory and discrete choice models

In chapter 2 the micro-economic approach for representing a spatial system was presented and discussed. Although the models that can be derived have a strong theoretical content, the nature of the simplifying assumptions, their mathematical representation and their operational difficulties, forces them to stay as theoretical propositions only, providing no useful analytical tools for the practitioner. The main shortcomings of this approach can be summarised as follows:

(a) Both users and suppliers are assumed to have perfect information about market conditions. Limited information would be a more realistic assumption.
(b) Bidders or suppliers are assumed to have deterministic utility functions. Faced with the same options, they will always make the same choices. Variations in individual behaviour should be considered instead.
(c) Bidders or suppliers have frictionless mobility, and they can even appear or disappear at no cost. Restricted dynamics would produce more realistic results.
(d) There is no way to aggregate demand or supply functions. With a large population, the resulting models are impractical. If, in order to solve this, the analyst applies the same models to socio-economic and spatial aggregates, the results would be unreal, because it would imply large groups showing identical behaviour.

Spatial interaction models, as explained in chapter 3, were derived by maximising an entropy function, subject to some known constraints. The resulting distributions, as has been pointed out, are the least prejudiced statements about the system being modelled, that is, they are the distributions that make the weakest assumptions about the system apart from what is known and reflected in the constraints. As Williams (1977b) remarks, the method can be viewed as a statistical aggregation procedure, and any variability in the behaviour of individual members of the system may be considered to arise from a variety of sources which are not made explicit. It is not generally 'within the spirit of the method to pin down the variation to any particular source, as this would, in effect, present additional assumptions or hypotheses that could be exploited in model formation ... The methodology, by its very nature, generates descriptive rather than causal models because the fundamental sources of variability are not subject to a causal interpretation.' (Williams, 1977b).

There are, however, three main reasons to seek such a causal interpretation. Firstly, if spatial interaction models are interpreted in causal (economic) terms, a better explanation of the system being modelled can be achieved and, particularly, a relationship between the behaviour of the system at the macro-level and the behaviour of individuals can be established. Secondly, if we accept that in the real system individuals will behave rationally, even if it is in a limited way, it will be possible to simulate the functioning of the various market mechanisms, such as land, floorspace, transport, services, and so on. In this way, the theoretical models derived from micro-economic theory can be integrated with aggregate spatial interaction models. Thirdly, if a restricted rational behaviour is assumed in the decision-making process which results in particular spatial structures, the utility that individuals perceive from these decisions can be made explicit. This means that if spatial interaction models can be interpreted in terms of costs, utilities and elasticities, consistent user-benefit measures can be identified and evaluated.

In the following sections, random utility theory and discrete choice models will be presented. I will begin by describing the broad theoretical principles and the derivation of a general model. Then the concept of decision chains is developed to show how a complex model can be built. Finally, three related issues are discussed: variable costs, elasticities, and hierarchies.

4.1 Decision theory: the individual case

In general terms, decision theory describes social processes as a set of decisions made by individuals. The main assumption is that individuals choose rationally between the options available to them, subject to a number of constraints such as income, cultural background, and so on. For instance, when a household decides to buy a house, it will have a number of options, varying in price, size, type and location, and a number of restrictions, mainly in the scope of houses on offer and the available budget of the household. Individuals will choose from the options available to them, the one that renders the highest degree of satisfaction.

Each individual assesses the potential degree of satisfaction or utility he perceives from each one of the available options. He then ranks them according to the degree of satisfaction or utility perceived in each case, and chooses the one that provides the greatest utility. Utility, on the other hand, is a subjective concept – its perception will vary from one individual to another, and from one choice to another.

Mathematically, utility can be represented as a mathematical function for a particular individual and option, containing variables that describe measurable attributes of each option. Faced with a particular set of options, an individual may be assumed to evaluate each one with the *same* utility function. The utility derived from each option can be defined as:

$$u^{sk} = U^s(X^k, S^s), \qquad k \in B^s \tag{4.1}$$

where u^{sk} represents the level of utility an individual s would perceive should he choose option k. U^s is a utility function that the individual s applies to *all* options of a similar kind available to him, that is, to all $k \in B^s$, the set of options. It is understood that B^s contains those options that the individual s can afford and that they are *substitutes*; that is, they are alternative options that perform similar functions. X^k represents the measurable attributes of option k, and S^s represents the socio-economic characteristics of individual s.

If it is further assumed that the socio-economic characteristics of individual s are implicitly represented in the utility function, then equation (4.1) simplifies into:

$$u^{sk} = U^s(X^k) \tag{4.2}$$

The selected option will be k if $U^s(X^k)$ is higher than that of any other option within set B^s, that is, if:

$$U^s(X^k) > U^s(X^q), \qquad \forall q \in B^s, q \neq k \tag{4.3}$$

Expression (4.3) yields a unique solution, and is roughly equivalent to the micro-economic model, and thus is subject to criticisms (a) to (d) at the beginning of this chapter. In particular, it is of little practical value, since it would be impossible to keep track of the utility function of each individual living in a city or a region, and also because the number of options can be very large. There is, then, a need for aggregation. Individuals must be grouped according to their socio-economic characteristics, and options must be made discrete.

4.2 Decision theory: the aggregate case

It was mentioned that, even in the individual case, utility cannot be a deterministic function, because the same individual may behave in a slightly different way each time. When the aggregate case is considered, where utility functions must now refer to a population of individuals choosing among groups of options, variations become too big to be ignored. Naturally, the smaller the group, the better, but the analyst must strike a balance between the benefits of working with a large number of relatively homogeneous groups and the difficulties of operation, calibration and data collection.

In the aggregate case, then, it is no longer plausible to assume an invariable utility function. It must be assumed, instead, that utility will vary within the group around a mean value. If the population is very homogeneous, then all members will tend to perceive utility in a similar way, and thus, dispersion with respect to the mean value will be small (never zero, not even if the population is equal to one). On the contrary, if the population is large, variations around the mean value will also be large.

There are multiple *sources of variability* as groups increase in size. The list of sources proposed by Domencich and McFadden (1975) can be compiled together with that of Williams (1977a) and the author's own as follows:

(a) Intra-individual variations. Individuals do not necessarily perceive utility the same way each time, and even if they did, they might not respond with the same action.

(b) Intra-option variations. Options do not always present themselves in the same way every day. Buses, for instance, might be late one day, early another. A bicycle ride might be very attractive for a short trip in the proper season, but not, say, on a rainy day.

(c) Individuals within a group do not necessarily share the same information about the attributes of each option.

(d) The utility function in the model itself can be mis-specified. Perhaps the main variables are well represented in the model, but there might be others of lesser importance that are not represented in the model and that account for variations in the behaviour of individuals within a group.

(e) The exact spatial location of individuals within an aggregate group/zone will vary, both with respect to the origin of an interaction as well as with respect to its destination. Consider the example of figure 4.1 where two individuals of the same group travel from zone i to zone j. For a deterministic model, the measurable attributes of the perceived utility function would be the same for both individuals. However, an individual travelling from point a to point d would pay less than another travelling from b to c.

(f) The precise position of individuals within their socio-economic group is another important source of variability. Thus, if a group has been defined, say, as people with personal income between 1000 and 3000, an individual with an income 1100 might behave quite differently to another with an income of 2900.

(g) The criteria for defining the groups are never fully satisfactory. If, for instance, income has been chosen to aggregate population into groups, other criteria must be ignored, such as level of education,

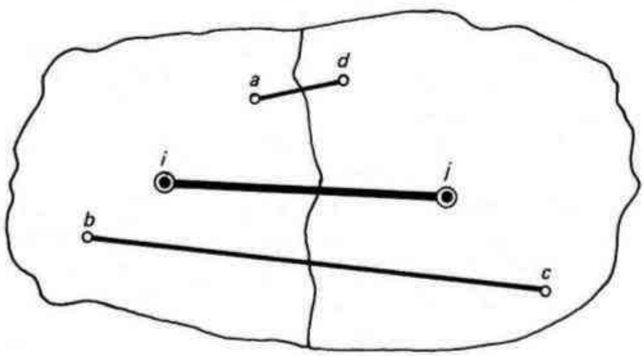


Figure 4.1. Variability in the cost of travel

cultural background, and so on. In the case of spatial aggregation, zone boundaries are always arbitrary.

(h) Individuals might not always select the option with the highest utility, even if they are aware of this, because they might be restricted by some other factor. For example, if an important change has been introduced in the transport system, many households might perceive that they are no longer living in the optimal place, but the cost of buying a new house and the cost of moving into it outweights the difference in the perceived utility of travel.

Aggregation, then, introduces sources of variability in the way utility is perceived by each individual member of the population. Thus, for the group as a whole, a distribution of perceived utilities will result. Figure 4.2(a) shows an example of the distribution of utilities of a hypothetical population with respect to a particular option. A large proportion of the individual members will assign to the option a level of utility of m in the diagram, which is the mean value for the population. Other members, however, will assign values smaller or bigger around the mean value. If the group is small and homogeneous, variations will be small also, resulting in the steep distribution of utilities A in figure 4.2(b). On the

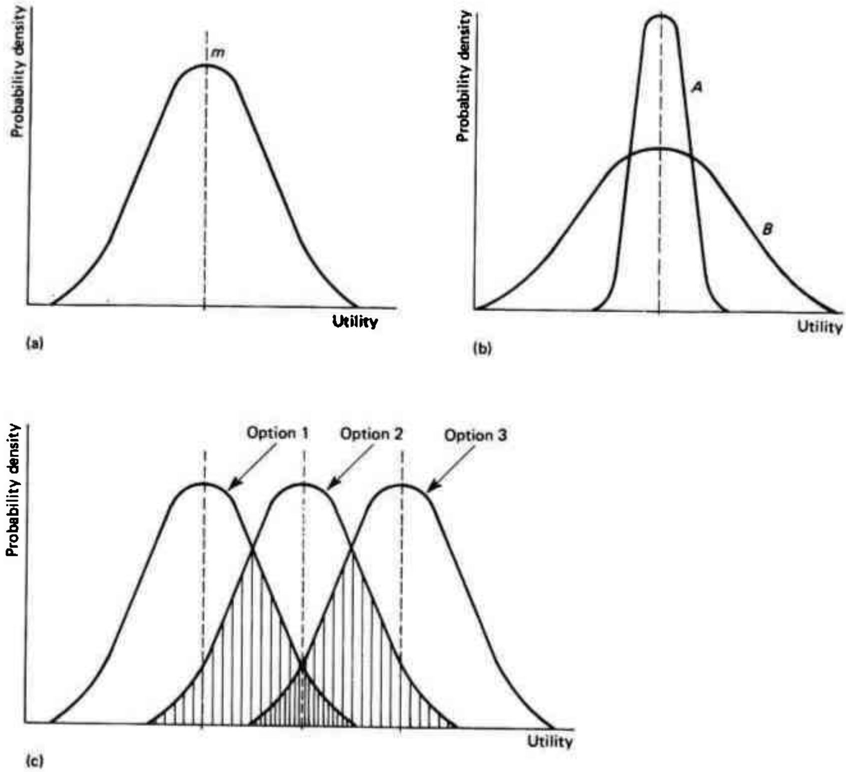


Figure 4.2. Distributions of perceived utilities

other hand, if the group is large and not homogeneous, the distribution of perceived utilities will look more like curve B in figure 4.2(b).

Figure 4.2(c) shows the result in the perception of utilities when the population is faced with three options. Since each individual member will judge each option with the same parameters, the result must be that the distribution of perceived utilities in the aggregate case will be identical in shape for the three options, shifting to the right or to the left according to the option. Those distributions further to the right will be, in general, preferred, but there are some important overlaps (shown with different shading in figure 4.2(c)) that explain why the population is distributed among all options, albeit in different proportions.

Mathematically, if equation (4.2) represents a deterministic utility function for an individual, an aggregated utility function for a population must include a random element:

$$u^{sk} = U^s(X^k, \zeta) \tag{4.4}$$

where ζ represents the random variation in the utility function. There will no longer be a unique solution: the introduction of sources of variability means that all options can be selected eventually by individual members of the group. A probability can then be assigned to each option. The probability P^{sk} that group s will select option k will be:

$$P^{sk} = \text{Prob}\,[U^s(X^k, \zeta) > U^s(X^q, \zeta)], \quad \forall\, q \in B^s,\, k \neq q \tag{4.5}$$

The random element ζ can be eliminated from the utility function by assuming that the function itself is random. Furthermore, the utility function can be divided into a deterministic function $V^s(X^k)$, representing its fixed components or measurable attributes or *strict utility*, and $\mu^s(X^k)$, a stochastic element: .

$$U^s(U^k) = V^s(X^k) + \mu^s(X^k) \tag{4.6}$$

which means that (4.5) can be transformed into:

$$P^{sk} = \text{Prob}\,[\mu^s(X^k) - \mu^s(X^q) < V^s(X^k) - V^s(X^q)], \quad \forall\, q \in B^s,\ k \neq q \tag{4.7}$$

If it is assumed that μ^s represents the degree of variation within the population but not within the options, then all μ^s can be said to have a common joint distribution. If $\tau(t^1, \ldots, t^N)$ is the cumulative joint distribution of $\mu^s(X^1), \ldots, \mu^s(X^N)$, and τ^k is the kth derivative of τ, then the probability that option k is selected is the integral of (4.7):

$$P^{sk} = \int_{-\infty}^{+\infty} \tau^k(t + V^{sk} - V^{s1}, \ldots, t + V^{sk} - V^{sN})dt \tag{4.8}$$

Equation (4.8) represents the basic model to simulate discrete decisions under the assumption of random utility. Many specific forms can be derived from this generic model, depending on the particular function chosen to represent the joint distribution τ.

Domencich and McFadden (1975) explored the major statistical functions that could be applied to represent τ to a considerable extent. The best-known functions are the normal, the logistic and the Cauchy distributions, which after integration yield the probit, the logit and the arctangent models respectively. Domencich and McFadden (1975) explored and tested extensively the three distributions, particularly for the most significant arguments (avoiding the extreme values, that is, close to zero or one). Considering two options, Domencich and McFadden argue that if $\mu^s(X^1)$ and $\mu^s(X^2)$ have a joint distribution, the distribution of the differences $\mu^s(X^1) - \mu^s(X^2)$ must also have the same form for the model to be consistent. From this point of view, if $\mu^s(X^1)$ and $\mu^s(X^2)$ are normal, then $\mu^s(X^1) - \mu^s(X^2)$ also has a normal distribution. Similarly, if $\mu^s(X^1)$ and $\mu^s(X^2)$ are independent Cauchy distributions, then $\mu^s(X^1) - \mu^s(X^2)$ also has a Cauchy distribution.

The logistic distribution does not have this property, but it can be solved by replacing it with a Weibull distribution, which also yields a logit model. The difference between a logistic and a Weibull distribution can be seen in figure 4.3. Apart from stability to the sum, already discussed, Weibull distributions have a property which is particularly relevant to this case: the distribution of the maxima of Weibull distributions is also a Weibull, which is very interesting for a utility maximising model. Based on this and other properties, the authors conclude that the choice of Weibull distributions must be preferred.

The question of whether the model should be a logit or a probit is still a matter of debate. Logit models are far simpler and easier to calibrate, but some authors (Daganzo 1980) argue that probit models produce more realistic results, because they are not affected by attribute correlations among options, making the additional effort worthwhile. This point will be discussed further below.

If the distribution is Weibull, the resulting multinomial logit model is:

$$P^{sk} = \frac{\exp(V^{sk} - \alpha^k)}{\Sigma_q \exp(V^{sq} - \alpha^q)} \tag{4.9}$$

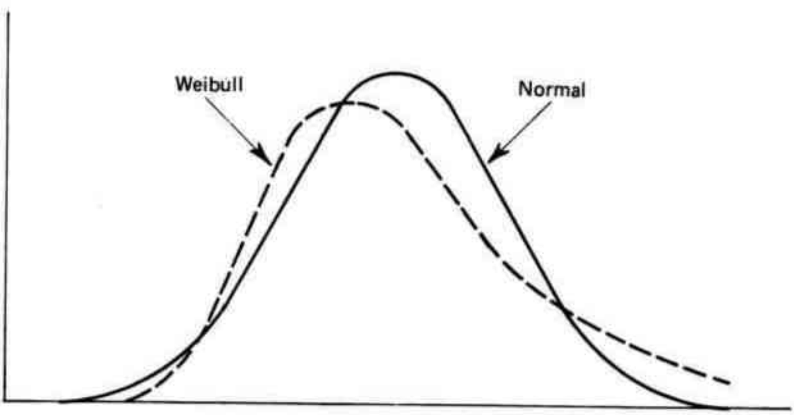


Figure 4.3. Normal and Weibull distributions compared

where α^k is a parameter which can be absorbed in the definition of V^{sk} without loss of generality. If, on top of this, an explicit parameter is introduced to the exponential function, the final form of the model is:

$$P^{sk} = \frac{\exp(\beta^s V^k)}{\Sigma_k \exp(\beta^s V^k)} \tag{4.10}$$

Cochrane (1975) makes a similar derivation of the model, but assumes from the start that the cumulative distribution of the joint probability τ can be approximated with an exponential function, as shown in figure 4.4, thus arriving, through integration, at the same multinomial logit model form. Cochrane, however, shows an important corollary to this, pointing out that if the model is of the form (4.10), then the average utility or benefit perceived after the population has been distributed to options is:

$$S^s = \frac{1}{\beta^s} \ln [\Sigma_k \exp(\beta^s V^{sk})] \tag{4.11}$$

where S^s is the average benefit perceived by group s, or *composite cost*, and β^s is the parameter of the exponential function. Furthermore, as Williams (1977a) points out, if two alternative policies, say 1 and 2, are compared, the difference in benefit, ΔS^s, is:

$$\Delta S^s = \frac{-1}{\beta^s} \ln \left[\frac{\Sigma_k \exp(\beta^s V^{sk}) 2}{\Sigma_k \exp(\beta^s V^{sk}) 1} \right] \tag{4.12}$$

Indicator ΔS^s is conceptually equivalent to the traditional consumer's surplus indicator, but helps to solve many problems related to evaluation. Williams (1977a) also makes a full review of the arguments of Domencich and McFadden, as well as those of Cochrane, adding a number of new insights to the derivation. Williams particularly emphasises the importance of indicators (4.11) and (4.12), pointing out that if it is accepted that (4.10) is the correct model to simulate decisions made by a population of individuals, then there are no degrees of freedom as to which is the user benefit indicator: it *must* be (4.11) and (4.12). Operationally, this formulation has the advantage of presenting a full integration between simulation and evaluation, and, as will be

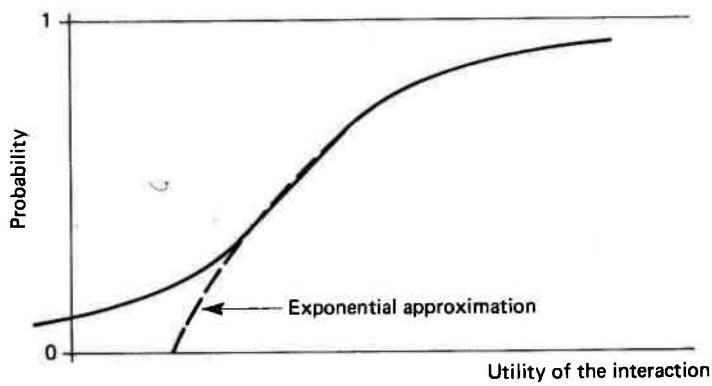


Figure 4.4. Cochrane's exponential approximation of the distribution of utilities

seen in the next section, can have an important effect in the results of a planning study.

4.3 Some properties of discrete choice models

To begin with, some properties of parameter β^s are worth noting. If, in a particular utility function, measurable attributes X^k represent costs, then the parameter will be negative. The value of the parameter is related also to the degree of dispersion of the distribution of utilities. If the parameter is large and negative, it indicates that all individuals within the group tend to agree on what is the best (least cost) option. The parameter is, hence, also related to the level of aggregation of the groups involved: if the decision-making group is small and homogeneous, the parameter will be large and negative. In the limit, when the group is restricted to one member, located in a zone so small that it becomes a point location, the parameter will be minus infinity, and the model will be equivalent to the deterministic micro-economic model. This is a very important consideration; in a residential location model, for instance, if the study area is divided into a small number of large zones, calibration will result in a relatively small negative value of β^s, and the opposite will result if the model is calibrated for a study area divided into a large number of smaller zones.

On the other hand, if there are several options k, one of them will represent the *minimum cost option*, say k^*. As the value of β^s tends to minus infinity, probability P^{sk^*} tends to one. However, as the value of β^s tends to zero, the probabilities of all options become equal, that is, $P^{sk} = 1/N^s$, where N^s is the number of options available to group s. To illustrate this point, consider the numerical example below, where a case with two options with costs 5 and 8 is calculated for two population types, represented in turn by two different parameters, β^s equal to -0.2 and -0.6 (Table 4.1).

It can be seen that, for the population with a low sensitivity to cost, that is, when the value of β^s is -0.2, the probability of the least cost option, when $V^{sk^*} = 5$, is 0.6457. By contrast, for the population with a high sensitivity to cost, that is, with a parameter of -0.6, the probability rises to 0.8581. In other words, 86% of the high sensitivity group will choose the least cost option. As a result, the average cost $\bar{c}$ paid by the high sensitivity group will be lower than that of the low sensitivity

Table 4.1.

V^{sk}	β^s	$\exp(\beta^s V^{sk})$	P^{sk}	$\bar{c}$	S^s (eq. (4.11))
5	-0.2	0.3679	0.6457		
8		0.2019	0.3543		
Total		0.5698	1.0000	6.0629	2.8126
5	-0.6	0.0498	0.8581		
8		0.0082	0.1419		
Total		0.0580	1.0000	5.4257	4.7450

group (5.43 against 6.06), because a higher proportion chooses the least cost option.

Table 4.1 also contains the computation of the average utility S^s with equation (4.11). In this case, the low sensitivity group perceives a lower dis-utility from the same choice set (2.81 against 4.75). Because this group is less sensitive to cost, it will be better off when confronted with cost-related options, even if it pays a higher average cost. In general, it can be seen that, as the absolute value of the parameter rises, so does the average dis-utility. In the limit, when β^s is minus infinity, the probability of the least cost option converges to 1.00 and the average utility will converge to the cost of the least cost option. Note also that if the choice set has only one option, equation (4.11) will always yield the cost of the single option, whatever the elasticity, with $\bar{c} = S^s$.

Assume, now, that a new, more expensive option is introduced to the system, with a cost of 12. Table 4.1 is changed as shown in table 4.2.

As can be seen, 14% of the low sensitivity group chooses the new option, against only 1% of the high sensitivity group. As a result, the average cost of the former group rises from 6.06 to 6.88 units, while the latter group only moves from 5.42 to 5.51 units. The average utility indicators show in both cases an improvement when the new option is introduced, but they also show that the less sensitive group benefits more, because the dis-utility moves from 2.81 to 2.07 while the high sensitivity group hardly moves from 4.75 to 4.72. Now equation (4.12) can be applied to quantify the consumers' surplus in each case:

$$\text{low sensitivity} \quad \Delta S^s = \frac{1}{-0.2} \ln \left[\frac{0.6605}{0.5698}\right] = -0.7386$$

$$\text{high sensitivity} \quad \Delta S^s = \frac{1}{-0.6} \ln \left[\frac{0.0588}{0.0580}\right] = -0.0228$$

The results above clearly show that the low sensitivity group will benefit more from the introduction of the new option, even if it will pay more money in the new situation. The numerical results have been calculated with program UTIL, but the reader can easily check that if a new cheaper option had been introduced, both groups would be much better off than if a more expensive option is introduced, and that in such

Table 4.2.

V^{sk}	β^s	$\exp(\beta^s V^{st})$	p^{st}	$\bar{c}$	S^s (eq. (4.11))
5	−0.2	0.3679	0.5570		
8		0.2019	0.3057		
12		0.0907	0.1373		
Total		0.6605	1.0000	6.8772	2.0738
5	−0.6	0.0498	0.8472		
8		0.0082	0.1401		
12		0.0008	0.0127		
Total		0.0588	1.0000	5.5092	4.7227

a case the high sensitivity group would benefit more. When calculating the total benefit after the new option has been introduced, the analyst must multiply the surplus of each group by the corresponding population and by the number of selections that each group will perform within the evaluation period. For instance, if the above example refers to a travel mode choice, and the question was to estimate the benefits of introducing a new, more expensive bus service during a year of operation, the calculation in table 4.3 should be performed.

These results also show the importance of indicators S^s and ΔS^s in evaluating policy options. Traditionally, transport related projects have been evaluated with a cost and time criterion, assuming that the preferred project will be the one producing the least average cost $\bar{c}$. The numerical example above, as well as the one that will be described below, show that this is clearly a fallacy. A new option will always produce benefits, however small, and these benefits will not be the same throughout population groups. It is surprising, then, that the $\bar{c}$ indicator is still used in so many studies!

The next numerical example, to illustrate these points further, was calculated with program RIVER, and can be seen in figure 4.5. The example assumes an initial situation 1, where a region has been divided into three zones. Zone 1 is separated from zones 2 and 3 by a river which cannot be crossed regularly. People living in zone 1, aware of their isolation, are putting pressure on the government, and as a consequence, the construction of a bridge is being considered with highway connections from zone 1 to both 2 and 3. A planner is brought in to determine if such a bridge should be built.

In order to assess this, the planner applies a simple residential model and calibrates it against the real situation 1. He then feeds the model with the hypothetical situation 2 and compares the results. The form of the residential model, assuming a single population group, is as follows:

$$T_{ij} = \frac{E_i L_j \exp(-\beta d_{ij})}{\Sigma_j L_j \exp(-\beta d_{ij})}$$

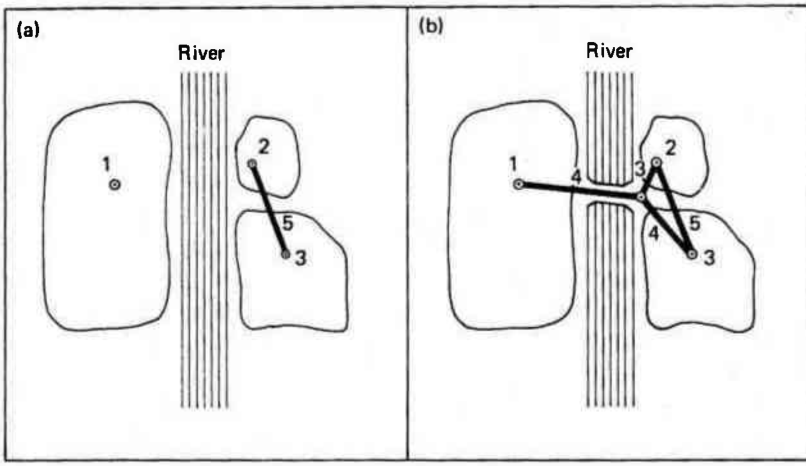
where T_{ij} is the number of employees working in i and living in j, E_i is the number of jobs in i, L_j denotes residential land available in j, d_{ij} is the distance between i and j, and β is the utility parameter.

The analyst begins by collecting the appropriate data: $E_i = (100, 200, 50)$; $L_j = (30, 10, 20)$; distances are computed from the map in

Table 4.3.

	Low sensitivity group	High sensitivity group
Population	1000	800
Trips per year	380	450
ΔS^s	−0.7386	−0.0228
$\Sigma_s(\Delta S^s$ Pop. trips)	−280668	−8208
Total annual surplus	288 876 units	

figure 4.5(a), with internal distances set to one and the distance between zone 1 and the rest set to infinity. Assume that calibration resulted in $\beta = 0.3$. The results show that the 100 employees working in zone 1 have no choice but to live in zone 1, and that there is a certain



(a)

	Empl. E_i	Land L_j
1	100	30
2	200	10
3	50	20

Distances c_{ij}	1	2	3
1	1	∞	∞
2	∞	1	5
3	∞	5	1

T_{ij} =	1	2	3
1	100	0	0
2	0	125	75
3	0	7	43
Resulting pop. Σ	100	132	118

	Average cost	Average utility (Eq. (4.11))
1	$\bar{c}_i$ = 1.00	S_i = 3.31
2	2.48	5.42
3	1.48	4.20
	$\bar{c}$ = 1.91	S = 4.64

(b)

	Empl. E_i	Land L_j
1	100	30
2	200	10
3	50	20

Distances c_{ij}	1	2	3
1	1	7	8
2	7	1	5
3	8	5	1

T_{ij} =	1	2	3
1	88	5	7
2	47	95	57
3	7	6	37
Resulting pop. Σ	142	107	101

	Average cost	Average utility (Eq. (4.11))
1	$\bar{c}_i$ = 1.79	S_i = 2.89
2	3.55	4.53
3	2.46	3.71
	$\bar{c}$ = 2.89	S = 3.94

User benefits

ΔS_i =		
1	42.72	
2	179.75	(Equation (4.12))
3	24.69	
ΔS =	247.16	

Figure 4.5. Example of the river to illustrate the way average costs and user benefit indicators should be calculated: (a) situation 1 ; (b) situation 2

proportion of people that work in zone 2 and live in zone 3 and vice versa. These results are represented in the T_{ij} matrix, and if the columns are added, the simulated number of residents can be computed ($\Sigma_i T_{ij}$). The resulting population is $R_j = (100, 132, 118)$, and if the model has been well calibrated, it should match real data.

The analyst now feeds the model with hypothetical information describing situation 2. The only change is that the bridge is now in operation, thus altering the c_{ij} matrix, replacing the infinite values by finite ones. Applying the same residential model, the analyst obtains a new T_{ij} matrix, with a different residential location. This time, people working in zone 1 will be able to live in other zones, improving their residential choice; those working in zones 2 and 3 now have an additional opportunity of living in zone 1, which they did not have before. The result of this, as shown in figure 4.5(b), is that the number of households living in each zone is now $R_j = (142, 107, 101)$; in other words, the population living in zones 2 and 3 has decreased, and the population living in the isolated part of town has increased, i.e. the town has expanded.

In order to determine which of the two situations is better, the analyst proceeds to calculate accessibility indicators, since this is the main purpose of the proposed bridge. The traditional mean travel cost is calculated as:

$$\bar{c}_i = \frac{\Sigma_j T_{ij} d_{ij}}{\Sigma_j T_{ij}} \quad \text{and} \quad \bar{c} = \frac{\Sigma_{ij} T_{ij} d_{ij}}{\Sigma_{ij} T_{ij}}$$

The analyst will find that the original pattern in situation 1 is $\bar{c}_i = (1, 2.48, 1.48)$ with an average of $\bar{c} = 1.91$, and that this deteriorates to $c_i = (1.79, 3.55, 2.46)$ with an average of $\bar{c} = 2.89$ after the bridge is built in situation 2. According to these results, the bridge should not be constructed, because people living in any of the three zones travel longer distances in the proposed situation. Hence, the planner rejects the project.

To the layman, the results are clearly absurd, because common sense dictates that the bridge is obviously an improvement, especially if the cost of the bridge is not being taken into consideration in the evaluation. The anomaly arises because accessibility is not being correctly measured. If, however, the planner had been educated in random utility theory, he would have used equation (4.11) instead, and would have obtained completely different results. Average utilities will change from (3.31, 5.42, 4.20) in situation 1 to (2.89, 4.53, 3.71) in situation 2 with weighted averages from 4.64 evolving to 3.94. Hence this planner would recommend the construction of the bridge. Furthermore, he can state that the populations in all three zones benefit, but those who work in zone 2 benefit more, by applying equation (4.12) which yields (42.72, 179.75, 24.69). Employees of zone 2 benefit the most because this is the largest concentration of jobs but with the smallest amount of residential land; the bridge offers them a new area for residential location which they did not have before. Employees of zone 3 benefit the least because they did have plenty of land to locate.

For once, planning decisions match the intuitive appreciation of those involved! What is even more striking is that the average cost criteria still dictates many decisions of this kind today; such criteria dominate current evaluation methodologies in many world banks and international development agencies involving huge amounts of money. Naturally, everyday planning problems are not as dramatic or as clear cut as the example above, but every exercise will carry some ingredient of this sort.

4.4 Decision chains

The paragraphs above have dealt with one particular choice. In an urban or regional system, however, a population group faces a large number of choices that are related to each other. In analytical terms, the set of choices can be represented in the form of a *decision chain.* In an urban context, a typical chain would be, for example:

place of work → residence → shopping → transport mode

Each link along the chain is clearly conditioned by the preceding link. For instance, where to go shopping is a decision conditioned by the place of residence. In order to represent such a decision chain in a set of models, each component must precede the next in the right order. If the place of work is a given starting point, each link along the chain can be represented by a corresponding model, producing probabilities such as $P(r)$, $P(s)$ and $P(m)$. The simulated decision chain will be:

place of work $\rightarrow P(r) \rightarrow P(s) \rightarrow P(m)$

The number of people that go shopping by bus from their place of residence given that they work in a particular zone can be calculated as the number of people that work in that zone $\times\ P(r) \times P(s) \times P(\text{bus})$. This is quite a comfortable solution, because it is possible to model each link of the decision chain separately through independent multinomial logit models, thus avoiding very large simultaneous computations.

For this to be possible, however, a particular condition must be fulfilled, which has been defined by Luce (1959) as the *independence of irrelevant alternatives* axiom. This axiom requires that the options being chosen be independent of the presence or absence of other non-chosen alternatives. In the example above, it means that the mode choice, for instance, is independent from other choices. Once the option to go shopping to a particular place has been selected, the choice of bus is a completely independent one.

However, the problem is more complex than this, because each link in the chain may influence the preceding ones. In the above example, it could well be that people decide where to go shopping precisely because there is a good bus service. Thus, the choice of transport mode affects the choice of shopping. Similarly, the choice of residential location may have been influenced by the availability of local shopping facilities.

In order to accommodate this, the process of calculation must begin from the other end, that is, from the last link in the chain, proceeding backwards. In the example, we would have to calculate the overall aggregate availability of transport from residence to shopping, that is, the composite cost of travel. Following Williams (1977a), the proper way to do this is by applying equation (4.11):

$$S^m = \frac{1}{\beta^m} \ln \left[\sum_k \exp\left(\beta^m V^k\right) \right] \tag{4.13}$$

where S^m is the aggregated travel utility, β^m is the parameter regulating mode choice, and V^k is the strict utility of each mode k, and where summation is over all modes available for the decision-making population and the particular origin–destination pair.

If we assume, for simplicity, that the strict utility of shopping choices is determined only by transport availability, then the aggregate utility of all shopping facilities for a particular residential choice will be:

$$S^s = \frac{1}{\beta^s} \ln \left[\sum_m \exp\left(\beta^s S^m\right) \right] \tag{4.14}$$

where S^s is the aggregate shopping utility, β^s is the parameter regulating the shopping place choice, and S^m is the strict utility of each shopping place, calculated in (4.13), and where summation is over all shopping places available for the decision-making population and the particular origin–destination pair.

Finally, if we assume, for simplicity, that the strict utility of residential choices is determined only by shopping availability, then the aggregate utility of residential locations, given a particular place of work will be:

$$S^r = \frac{1}{\beta^r} \ln \left[\sum_s \exp\left(\beta^r S^s\right) \right] \tag{4.15}$$

where S^r is the aggregate residential utility, β^r is the parameter regulating the residential choice, and S^s is the strict utility of each residential place, calculated in (4.14), and where summation is over all residential places available for the decision-making population and the particular origin–destination pair.

In this way the transport element has been transferred into the estimation of the shopping utility, and, eventually, into the residential utility. On reaching the top of the decision chain, the calculation sequence must reverse direction and follow the chain again in the original direction, calculating the probabilities:

$$\begin{array}{ccccccc} \text{place of work} & \rightarrow & P(r) & \rightarrow & P(s) & \rightarrow & P(m) \\ & & \uparrow & & & & \\ & & S^r & \leftarrow & S^s & \leftarrow & S^m \leftarrow \text{start here} \end{array}$$

This particular way of arranging the set of models and aggregating utilities is called *nested multinomial logit models* (NML). It is a particularly powerful structure for the representation of complex urban or regional systems.

4.5 Variable costs and elasticities

The calculation process described above begins by the aggregation of utilities or composite costs following the decision chain backwards, and then calculating probabilities forwards. The calculation process would end here were it not for *variable costs* and *elasticities*.

Variable costs emerge because some, or even all, links in the chain may have capacity problems and consequently vary their costs. In the above example, if the bus service is used beyond its capacity, the cost of travel (or time) may increase until residents eventually choose other options. This effect will be transferred to the shopping place and residential choices. Shopping places, in turn, may only cater for a limited number of customers, and these might want to avoid overcrowding. In the case of shopping, however, it could happen that the prices of commodities being sold decrease due to economies of scale. Residential areas will also have a limited capacity, and as demand increases for a particular location, so will the price of land. Transport costs, retail prices and the price of land should be included in the representation of strict utilities.

Demand elasticities also influence the process; in the example, if bus services to shopping facilities get congested, people might travel less, say once a week instead of once a day, thus generating fewer trips. In the case of shopping and residential location it is more difficult to establish the presence of elasticities.

In order to represent variable costs and elasticities, the calculation process must become iterative. First, utilities are aggregated backwards along the chain and probabilities are estimated forwards in the usual way. Once demand has been assigned to all options, costs must be adjusted according to the relationship between demand and supply. If the cost of each option has been modified, the calculation of aggregated utilities must be re-evaluated in a new iteration, thus starting the process all over again. In a second iteration, the calculation of probabilities will change and so will the magnitude of the demand to be assigned. This is repeated several times until a state of convergency is reached. Convergency can be measured as the degree of change in costs between one iteration and its preceding one; if the degree of change is less than a pre-established convergency criterion, then calculation ends. The calculation process must be modified, then, as follows:

$$\begin{tikzcd}
\text{place of work} \arrow[r] & P(r) \arrow[r] & P(s) \arrow[r] & P(m) \arrow[d] & \\
& & & \text{adjust costs} \arrow[d] & \\
& & & \text{evaluate convergency} \arrow[r] \arrow[d] & \text{exit} \\
& S^r \arrow[uuu] & S^s \arrow[l] & S^m \arrow[l] & \text{start here} \arrow[l]
\end{tikzcd}$$

4.6 Hierarchies

Sometimes, for analytical purposes, a particular choice can be divided into sub-choices, and this gives rise to the concept of *hierarchies* in the decision-making process. For example, residential choice could be divided into two hierarchical levels: first the choice of a district within the city, and then the choice of a neighbourhood within a district. Similarly, a hierarchy of transport modes can be established: first the choice between public and private transport, and then the choice between buses and other options such as trains or subways.

These hierarchical levels can be represented and related to each other in exactly the same way as different links in the decision-making chain, and the computing process is identical. Note that because a given decision is split into hierarchical levels, the resulting β parameters are bound to be different.

The use of hierarchies can be very useful because it permits some parts of the system to be represented in great detail, while other parts are dealt with in less detail. For instance, it may be of interest to investigate residential location within a particular district: in this case the residential model can determine the probable location of residents in all districts at a first hierarchical level, and apply a second level of analysis only to the district in question. In this way the analyst saves in data collection and computation.

The concept of hierarchies must be taken into consideration even if it is not used explicitly. For instance, when an urban area is divided into zones for analytical purposes, it is important that the resulting zones are relatively homogeneous, and that the levels of analysis are not confused. It would be an error, for instance, if the area is divided into an heterogeneous set of districts and neighbourhoods at the same hierarchical level. After calibration, the resulting β parameter, because it is linked to the level of aggregation, will fit neither of the two sets of zones properly.

The use of hierarchies also has a theoretical importance. It was mentioned that multinomial logit models are bound to produce errors if there are attribute correlations between some of the choices being simulated. The multinomial probit model does not have this problem, but it is difficult to apply because of its complexity, and difficult to calibrate. Hierarchies provide the analyst with a way round the problem of attribute correlation when using multinomial logit models: those options within the choice set that are highly correlated can be grouped together to conform an aggregate option, and then, at a lower hierarchical level, treated as a separate choice set conditioned to the previous one. Examples of this will be seen in chapters 6 and 7.

The random utility model in its ML form is almost identical to the entropy maximising model of chapter 3. This is no coincidence, because both derivations arrived at the same solution from opposite ends. The maximum entropy approach assumes that choices are perfectly random, and then introduces a rational (cost) restriction. By contrast, random utility theory begins by assuming that choices are perfectly rational, and then, because of aggregation, introduces random elements.

5

Spatial accounting models

In the preceding chapters, micro-economic theories of the use of space were reviewed, followed by spatial interaction and entropy maximising models, and finally random utility theory was described. This chapter concludes the general theoretical framework by describing another important development from macro-economics: the input–output model. This introduces a new dimension to the problem, that of production and its relation to the urban and regional structure.

In very broad terms, the first generation of input–output models were intended as nationwide global economic accounting frameworks; the second generation of I–O models attempted a regional disaggregation; a third generation, currently in progress, attempts a more general description of the structure of a nation, and consequently these models have been defined as social accounting models. The intention of this chapter is to derive a general model to represent a spatial-economic system at any scale, capitalising on the teachings of random utility theory. Since spatial aspects become so closely linked to the economic accounting ones, the results have been termed *spatial accounting models.*

Keynes' theory of production and the multiplier is briefly reviewed as a starting point for later developments of regional theories such as the theory of regional rent and the study of the economic base of a region. Then the elements of the input–output model are presented and its regional disaggregation. At this point, the presentation diverts from the main stream by re-introducing random utility concepts, with related decision chains, variable costs, elasticities and hierarchies.

5.1 The theory of production and the multiplier

Keynes (1936) begins by introducing the principle of *effective demand,* whereby the process of production is mainly determined by consumption. According to this principle, producers determine the level of their activity according to their estimates of future demand. But demand is not considered as a function of prices, as in the classical approach, but rather of income, which in turn depends on investment. Formally:

$$P\{C[Y(I)]\} \tag{5.1}$$

where P represents production, C represents future demand or consumption, Y represents income, and I represents investment.

In the economic system as a whole, two main groups of interrelated actors are distinguished: producers and consumers. Producers can either produce or invest for future production. Consumers can either consume or save for future consumption. As shown in figure 5.1, producers pay consumers in the form of wages, profits, and so on. In turn, consumers pay producers in the form of consumption.

National product is defined as the sum of all goods produced, plus investment. National income is defined as the sum of consumption plus savings. Since the system is assumed to be closed, all goods produced must be consumed, so that a general equilibrium condition can be established:

$$Y = P \tag{5.2}$$

Since by definition $Y = C + S$ and $P = C + I$, it follows that the general equilibrium condition can be restated as:

$$S = I \tag{5.3}$$

that is, that in equilibrium savings must equal investment. Should consumers decide to save a larger proportion of their income, they would consume less, and consequently producers would be forced to produce less. This, in turn, reduces the income of the consumers via wages and profits, so that they will find it more difficult to save.

From the point of view of the consumers, the proportion of their income that they are willing to spend is called the *propensity to consume*, which is represented graphically in the consumption and saving schedules of figure 5.2. In figure 5.2(a) curve *CC* is the propensity to consume schedule, and *SS* in figure 5.2(b) is the propensity to save schedule. Point *B* represents the *break-even* point, where all income is devoted to consumption; thereafter, as income *Y* increases, a certain proportion of income becomes available for saving in increasing quan-

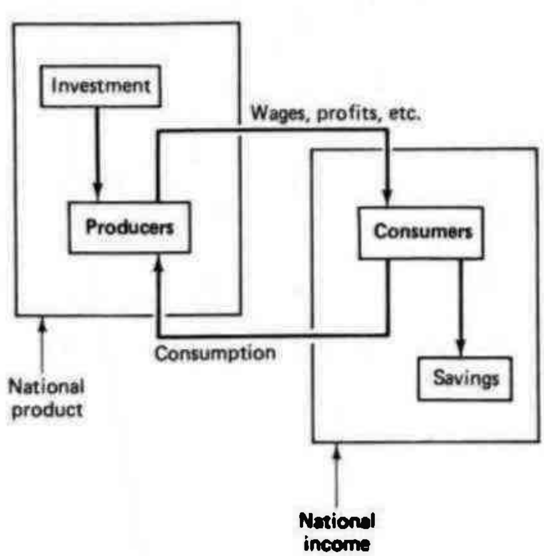
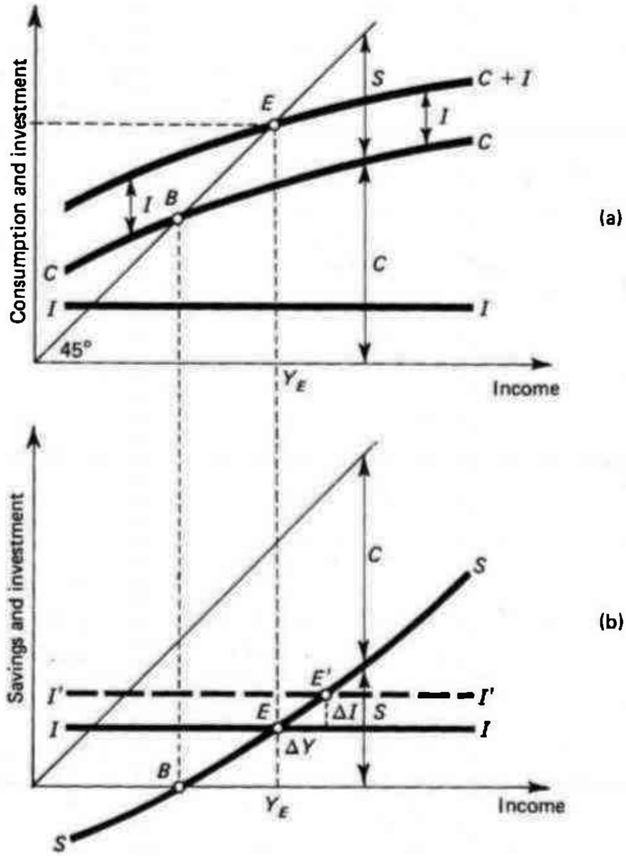


Figure 5.1. The general economic system in Keynes' model

Figure 5.2. Consumption and saving schedules in Keynes' model



tities, until it becomes equal to a given amount of investment represented by the horizontal line *II* (investment does not depend on income). When curve *SS* intersects *II* at point *E*, it means that savings equal investment, and hence equilibrium has been reached because of equation (5.2).

If for any reason there is an increase in savings, curve *SS* will shift upwards and point *E* will move to the left, thus causing the income at equilibrium Y_E to decrease. If, on the other hand, investment increases (because producers are confident of future economic prospects or because of state intervention), point *E* will shift to the right, such as *E′* in figure 5.2, and thus a higher level of income can be achieved. If investment increases, so will income, but in a larger proportion, which is due to the *multiplier* effects – a key element in Keynes' theory. The multiplier, *k*, expresses the effect on income produced by a unit increase in investment. That is:

$$Y = kI \tag{5.4}$$

or

$$k = \frac{dY}{dI} \tag{5.5}$$

from (5.2) and (5.3), and since $C = C(Y)$:

$$I = Y - C(Y) \tag{5.6}$$

Differentiating with respect to I and re-arranging the terms:

$$\frac{dY}{dI} = \frac{1}{1 - dC/dY} \tag{5.7}$$

As can be deduced from figure 5.2(b) dC/dY is the marginal propensity to consume, c. Correspondingly, $dS/dY = s$, the marginal propensity to save. Since according to (5.5), $dY/dI = k$, equation (5.7) can be reduced to:

$$k = \frac{1}{1 - c} \tag{5.8}$$

Since $C = Y - S$, equation (5.8) can also be re-written as:

$$k = \frac{1}{1 - (1 - dS/dY)} = \frac{1}{dS/dY} = \frac{1}{s} \tag{5.9}$$

The multiplier effect can be seen graphically in figure 5.2(b). If investment increases for some external reason from I to I', ΔY would represent the increase in income as a result of the increase in investment ΔI. Because SS tends to be below the 45° line, ΔY will tend to be larger than ΔI, thus resulting in a net gain in the system.

Keynes' theory is a powerful argument in favour of deficit spending through public investment, fiscal policies to stimulate consumption and investment, and as such, has had a profound social, economical and political impact. From a theoretical point of view, it is the basis for many later developments, some of which will be presented in the next sections.

5.2 The regional income model and the economic base method

Keynes' model was designed for applications at a national scale. The regional income (RI) model draws heavily from Keynes' model, but was designed for applications to a region within a country, and consequently achieves a greater degree of disaggregation by explicitly representing external factors, such as imports and exports, and distinguishing between private and public sectors. Keynes' model assumes a closed economy; the main intention of the RI model, by contrast, is to represent a regional economy within a larger context. In this case trade with other regions, particularly exports, are considered to be the main lever of development, instead of investment as in the Keynesian model.

The RI model begins with the following relationship:

$$Y = P = C + G + X - M \tag{5.10}$$

where C is private consumption, I is investment, G is government consumption, X represents exports, and M represents imports. Consumption, investment and imports depend on the level of income of the region, whereas exports depend on the income of other regions, and consequently, are exogenous to the model.

Figure 5.3 represents the main relationships between these variables. Curve E is the result of adding consumption, investment and government expenditure, and thus increases with income. Curve E' represents all aggregate demand for the production in the region: $E' = E + X - M$. Exports are assumed to be constant, since they do not depend on local income. Since $Y = P$, equilibrium is reached at point Q, where $Y = E'$.

Assume now that a higher level of exports can be achieved, say from X to X' (ΔX). A new aggregate demand will emerge, represented by E''. The equilibrium point will then shift from Q to T, and because of the multiplier effect, ΔY will tend to be larger than ΔX. The magnitude of the multiplier in this model will depend on the *marginal propensity to import*, and on the *marginal propensity to consume*.

The level of E can be considered as a function of income:

$$E = a + bY \tag{5.11}$$

where a is a constant equal to the level of E when Y is zero (point a in figure 5.3), and where b is the marginal propensity to consume:

$$b = \frac{dE}{dY} \tag{5.12}$$

On the other hand, imports are also a function of income:

$$M = f + gY \tag{5.13}$$

where f is a constant equal to the level of imports when Y is zero (point f in figure 5.3), and where g is the marginal propensity to import:

$$g = \frac{dM}{dY} \tag{5.14}$$

If $Y = E + X - M$, then by replacing E and M from (5.11) and (5.13):

$$Y = a + bY + X - f - gY \tag{5.15}$$

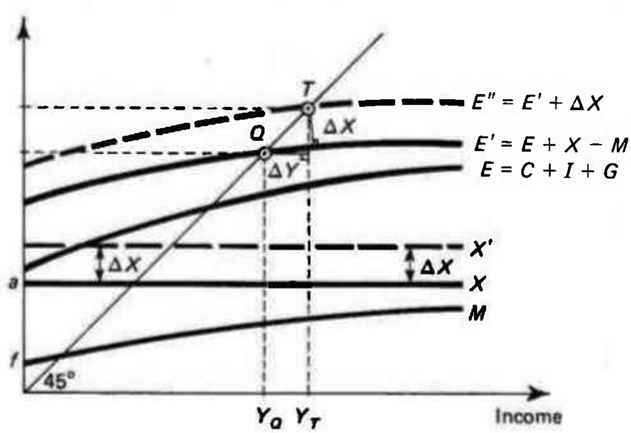


Figure 5.3. Basic relationships in the regional income model

from where

$$Y = \frac{X + a - f}{1 - (b - g)} \tag{5.16}$$

$$\frac{dY}{dX} = \frac{1}{1 - (b - g)} \tag{5.17}$$

As can be seen, in this model the multiplier is a function of the difference between the marginal propensity to consume and to import, while Keynes only considered the former.

Although the RI model is meant to be an operational model, it is difficult to apply, because the necessary information is generally not available. This led to the method or study of the *economic base* of a region, which is a more pragmatic approach and uses information readily available.

The main assumption of this method is that, following the RI model, the income of a region depends on those productive activities that can export goods or services to other regions. A key element in the method is to find ways of distinguishing between those activities oriented towards exports, or *basic sectors*, and those activities oriented to internal consumption, or *non-basic sectors*. Several methods have been proposed to make such distinction (see, for instance, Isard, 1960).

Once employment has been classified into basic and non-basic, a further simplifying assumption is introduced: that regional income will be proportional to basic employment. It is also assumed that the marginal propensity to consume is equal for all regions. Under these assumptions, the multiplier can be estimated as:

$$\frac{dY}{dX} = \frac{1}{1 - (E^s/E^t)} \tag{5.18}$$

where E^s represents non-basic employment (service) and E^t is total employment. Since E^t is the sum of all basic and non-basic ($E^t = E^b + E^s$), then:

$$\frac{dY}{dX} = \frac{E^t}{E^b} \tag{5.19}$$

5.3 The input–output model

5.3.1 The single-region input–output model

The input–output (I–O) model, introduced by Leontief (1941), represents a major attempt at disaggregating Keynes' model and the RI model by economic sector, thus providing a detailed account of the multiplier effect. It provides a general accounting framework for the representation of an economic system.

Figure 5.4 represents the main elements of the input–output table for a single region. The table makes a first distinction between inputs as origins of *producing sectors*, and outputs or *purchasing sectors*. In turn,

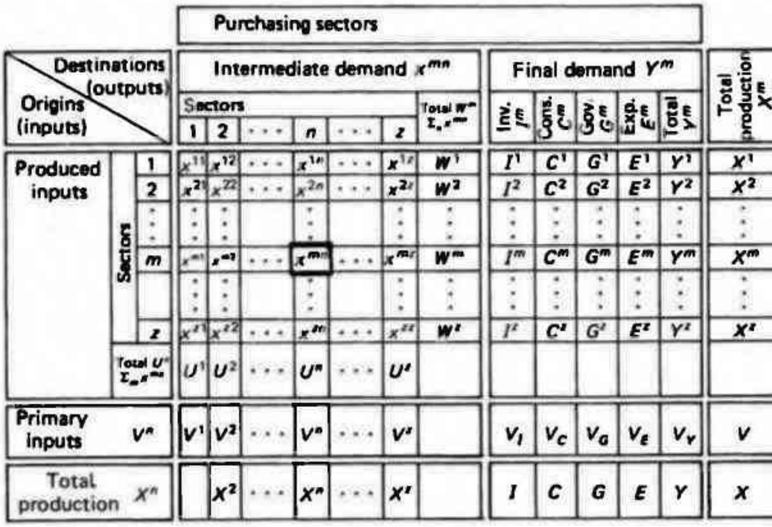


Figure 5.4. Basic elements of the single-region input–output accounting table

inputs are split between *produced* and *primary* inputs, and the purchasing sectors are divided into *intermediate* and *final demand* sectors.

Intermediate demand can be represented in a matrix x, in which each element x^{mn} represents the amount of output of sector m required for the production of sector n. The final demand sectors represent the end use given to what is produced in each sector m. The following sectors are normally distinguished within the final demand: I = investment, C = consumption, G = government, E = exports.

Produced inputs correspond to the origins of matrix x, and hence represent those inputs that are produced within the economic system. Primary inputs describe what was not produced within the system, and include elements such as labour, natural resources, imports, and so on.

A set of formal relationships can be readily established. Firstly, total production, X^m, in each sector m must be equal to the sum of all intermediate demands, plus all elements of the final demand:

$$X^m = \sum_n X^{mn} + I^m + C^m + G^m + E^m \tag{5.20}$$

Secondly, if W^m denotes total intermediate demand ($W^m = \Sigma_n x^{mn}$), and Y^m denotes total final demand ($Y^m = I^m + C^m + G^m + E^m$), then:

$$X^m = \sum_n x^{mn} + Y^m = W^m + Y^m \tag{5.21}$$

Thirdly, since total supply is assumed to be equal to total demand, total production in each sector must also equal the total value of inputs purchased from all sectors, U^n, plus the primary inputs, V^n:

$$X^n = \sum_m x^{mn} + V^n \tag{5.22}$$

Finally, a simplifying assumption is introduced: that there is a linear relationship between total production in any one sector, and the amount of production required fro.n other sectors. That is:

$$x^{mn} = a^{mn} X^n \tag{5.23}$$

where a^{mn} is a matrix of constants called *technical coefficients*, representing the amount of produce of sector m required to produce one unit of produce of sector n. Hence:

$$a^{mn} = \frac{x^{mn}}{X^n} \tag{5.24}$$

The set of technical coefficients a^{mn} is assumed to be stable in the short run, but could change in the long run due to changes in the technology of production. Thus, they are considered exogenous to the model. No economies of scale are assumed in these coefficients. Equation (5.21) can be re-written as:

$$X^m = \sum_n a^{mn} X^n + Y^m \tag{5.25}$$

Equation (5.25), the final form of the I–O model, represents a set of simultaneous equations in X^m that can be solved for a given set of final demands Y^m and technical coefficients a^{mn}. Once the system has been solved, the resulting set of X^m represents the amount of produce that is required to satisfy demand. At the same time, the resulting X^m represents a new demand that must be satisfied by further production, thus generating a fresh set of equations (5.25). Calculation then proceeds iteratively until final convergence is reached. This iterative process is shown in figure 5.5.

An alternative to the iterative solution can be achieved by matrix methods, which is the solution to be found in most textbooks, in the following way:

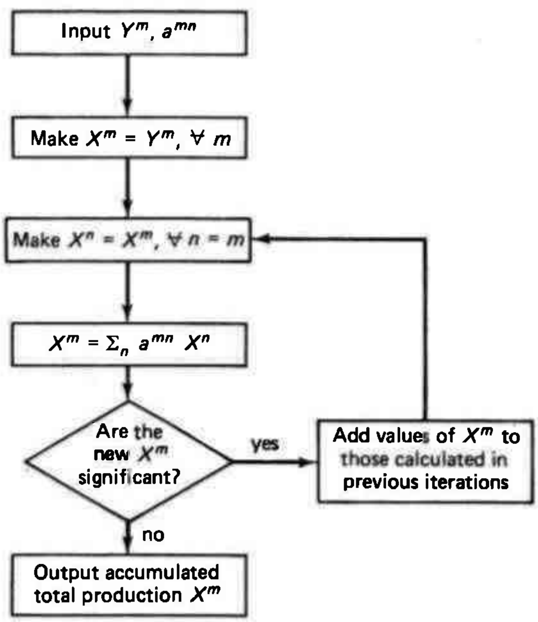


Figure 5.5. Algorithm for the solution of the single-region input–output model

$$X = (I - a)^{-1} Y \tag{5.26}$$

where in this case I is the unit matrix. This does provide a direct solution, replacing the need for iterations. However, it is easy to demonstrate that the interative solution in figure 5.5 is much more efficient than the inverse matrix solution of (5.26), a fact that recalls the advantages of algorithmic solutions discussed in section 3.5. This can be checked with program INPRO. Furthermore, the algorithmic solution converges in all cases to any desired degree accuracy, while the matrix method does not.

5.3.2 The multi-region input–output model

The most important problem when trying to apply the input–output model to a multi-region economic system is how to incorporate spatial interactions that occur between regions. The problem is far from trivial; it is not just a matter of adding a set of indices i and j to denote regions in the equations. Were we to do so, the main accounting relationship (5.25) would become:

$$X_i^m = \sum_n a^{mn} \sum_j X_{ij}^n Y_i^m \tag{5.27}$$

where X_{ij}^n is the flow of commodities n from region i to region j. Thus, $\Sigma_j X_{ij}^n$ represents the total amount produced in region i. Equation (5.27), however, does not suffice to solve the multi-regional system, because there is only one such equation for each region and sector ($i \times m$ equations), but many more unknowns ($i^2 \times m$) because of X_{ij}^n. What is required, then, is a set of equations in X_{ij}^n relating regions i to j.

Leontief and Strout (1963) solved the problem by assuming that all goods produced in region i go to a *supply pool*, and that all goods that are consumed in that region come from a *demand pool*. The authors then introduce a set of equations of the type:

$$X_{ij}^n = \frac{\Sigma_j X_{ij}^n \Sigma_i X_{ij}^n}{\Sigma_i \Sigma_j X_{ij}^n} Q_{ij}^n, \quad \text{for all } i \neq j \tag{5.28}$$

where Q_{ij}^n is some sort of distance function to be estimated from base year data.

The main relationships of this system can be seen in figure 5.6. It can be seen that a particular flow X_{ij}^n will represent the movement of commodity n from the producing region i to a supply pool. The supply pool for region i, together with the contributions of all other regions, make up the demand pool for region i, from where production flows back to any sector of region i (final demand, other sectors or the same sector).

From this starting point, Wilson (1970) attempts an integration between the Leontief–Strout proposition and spatial interaction models using entropy maximising principles, introducing the appropriate constraints. Wilson thus explores four types of model, from an uncon-

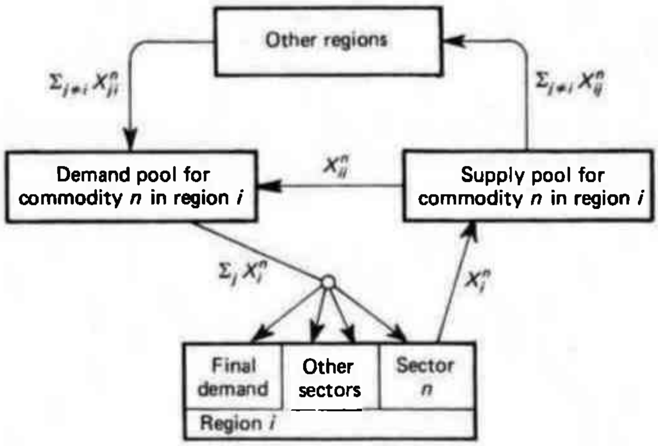


Figure 5.6. Flows of commodities in the Leontief–Strout multi-region input–output model

strained to a doubly-constrained gravity input–output model. The author then derives a hybrid model for different commodity types.

> It is clear that different types of goods will fall into different categories. For example, a primary commodity such as coal, is likely to be production-constrained; a commodity which is mainly an intermediate good for primary sectors would be attraction-constrained; a primary commodity which is an input for other primary sectors would be production–attraction-constrained; and there is a wide variety of goods which are neither production- nor attraction-constrained.
> (Wilson, 1970)

5.3.3 Concluding comments about input–output models

The single-region input–output model is a very powerful way of representing an economic system, and hence has become common practice in nationwide macro-economic planning. At such a scale, the information required to calibrate the model is generally available.

The need to disaggregate the model by regions stems from two main considerations. On the one hand, it enables analysts to assess the impact of economic policies at a regional level. For example, it can estimate the effects in all regions of an increase in the demand for steel in a particular region. On the other hand, it is quite clear that the way in which production is organised in the regional structure has an effect on the economy as a whole. Hence, a model that makes an explicit representation of regions and the way they interact is a better representation of an economic system than one that does not, even at a national scale.

Although many countries do collect the necessary data for Leontief and Strout's multi-region input–output model, such collection is costly and tends to produce unreliable results. Furthermore, the demand and supply pools scheme is cumbersome from a theoretical point of view. Wilson's proposition maintains the same criteria, and even if it does contribute some interesting insights to the problem, the resulting model is very complicated, and has not been used in practice.

What is required, then, is a general model based on a more consistent theoretical framework and that can facilitate implementation. Taking advantage of random utility principles, such a proposition is presented and discussed in the next section.

5.4 A random utility based multi-region input–output model

5.4.1 The general model

This proposition departs from the previous ones in that, instead of representing the flows as X_{ij}^{n}, a matrix X_{ij}^{mn} is considered as the main purpose of the simulation, whereby commodities flow directly from origin to destination regions and sectors. This scheme is represented in figure 5.7, where an equivalent to figure 5.6 has also been included.

In each region i and sector m there is a given amount of final demand, Y_i^m, which must be satisfied by producers. Potentially, all regions and sectors can produce goods to satisfy such a demand. The amount of goods n that are required to satisfy sector m in region i can be calculated by multiplying Y_i^m by the appropriate a^{mn} technical coefficients.

Demand activities are interested in acquiring the $Y_i^m a^{mn}$ goods at the best possible price, irrespective of where these goods come from. For the producers, the costs of producing the goods and of transporting them to where demand is located are the main factors influencing the prices of the goods they can supply. The difference between the cost of production plus transport and the price of the commodity in particular regions represents the surplus for producers, which they will try to maximise.

Put in these terms, each demand unit will assign its purchases to the producer that can supply at the lowest price. In other words, a demand unit will rank its supply options and will choose to buy from the winning supplier. The transactions will be:

$$X_{ij}^{mn} = Y_i^m a^{mn} min(K_j^n + c_{ij}^n + e_j^n) \tag{5.29}$$

where X_{ij}^{mn} represents the amount of produce that the demand unit of sector m in region i buys from a producer of n in region j (i and j could be the same region). K_j^n is the cost of producing n at j, c_{ij}^n is the cost of

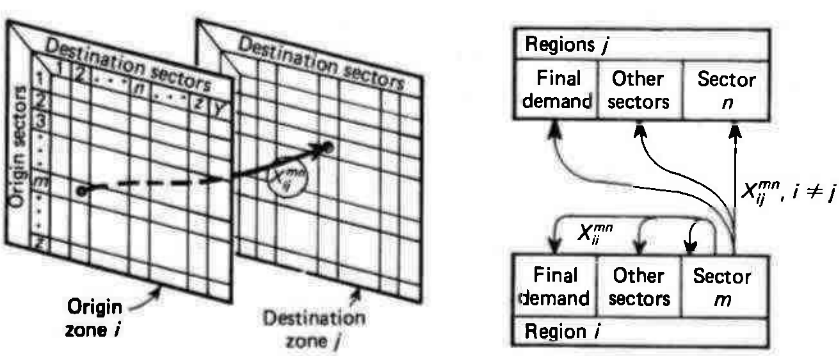


Figure 5.7. Multi-region multi-sector flow of commodities

transporting it to i, and e_j^n is the surplus for the producer of n at j. The minimum must be evaluated over all supply possibilities of n. Note that the surplus of the winning producer cannot exceed the difference between his own costs and those of the next producers down the list.

According to equation (5.29), the demand unit will buy all of its required goods n from the winning producer, and nothing from the rest. This would indeed be the case in a transaction between an individual demand unit and an individual producing unit. In turn, the producing unit will require goods from other sectors, and will assign its purchases in a similar fashion. This will generate a second link in a long decision chain. The links will branch out in tree fashion through individual producers of particular sectors and locations. Each branch will end when the amount being purchased becomes meaningless, when it encounters a primary input or when the purchase goes to a producer outside the system (imports).

However, as was discussed in the previous chapter, the individual case is of little practical value. Individual demand or supply units must be aggregated into sectors, and their point locations into regions or zones. The aggregation process, as was explained, introduces sources of variability into the system, such that the decision function must include random elements. The latter, as will be recalled, leads to a multinomial logit model. If the sum $K_j^n + c_{ij}^n + e_j^n$ is denoted by V_{ij}^n, then the distribution of purchases from sector n in i will become:

$$ {}^1X_{ij}^{mn} = Y_i^m a^{mn} \frac{\exp(-\beta^n V_{ij}^n)}{\Sigma_j \exp(-\beta^n V_{ij}^n)} \tag{5.30} $$

where ${}^1X_{ij}^{mn}$ represents the total amount of commodities of sector n purchased in j by the final demand of sector m located in region i. It is the first link in the production chain. Once equation (5.30) has been applied to all final demand sectors in all regions, it is possible to calculate the total amount of goods purchased in any one sector and region:

$$ {}^1X_j^n = \sum_i \sum_m {}^1X_{ij}^{mn} \tag{5.31} $$

In turn, in order to produce this first amount in each sector and region, the producing sectors must become demand sectors, generating further transactions:

$$ {}^2X_{ij}^{mn} = {}^1X_i^m a^{mn} \frac{\exp(-\beta^n V_{ij}^n)}{\Sigma_j \exp(-\beta^n V_{ij}^n)} \tag{5.32} $$

after which the total amount produced in each region and sector will be:

$$ {}^2X_j^n = \sum^i \sum_m {}^2X_{ij}^{mn} \tag{5.33} $$

For a generic link r:

$$ {}^rX_{ij}^{mn} = {}^{r-1}X_i^m a^{mn} \frac{\exp(-\beta^n V_{ij}^n)}{\Sigma_j \exp(-\beta^n V_{ij}^n)} \tag{5.34} $$

and

$$^{r}X_j^n = \sum_i \sum_m {}^{r}X_{ij}^{mn} \tag{5.35}$$

The total amount produced in a region and sector, once the system has converged, will be:

$$^{*}X_j^n = Y_j^n + {}^{1}X_j^n + {}^{2}X_j^n + \ldots \tag{5.36}$$

or

$$^{*}X_j^n = Y_j^n + \sum_r {}^{r}X_j^n \tag{5.37}$$

If the system is a rational one and transactions are expressed in money units, then a necessary condition is:

$$^{1}X_j^n > {}^{2}X_j^n > {}^{3}X_j^n > \ldots \tag{5.38}$$

that is, $^{r}X_j^n$ tends to zero as r tends to infinity. Program REGINP allows the user to explore this model.

5.4.2 Price accumulation in the production chain – normal profits

Following the decision chain from 1 to R, R being the last link, purchases are distributed among producers. Commodities move physically from the place of production to the place of consumption, and payments follow an opposite direction. If the chain is followed from R to 1, where 1 represents final demand, prices can be accumulated; a producing sector will charge a consuming sector an amount equal to its own production cost, plus transport cost plus profits; the consuming sector, in turn, buys commodities from several sectors and regions, and adds its own production cost and profits. When sectors have been aggregated, it is difficult to make a detailed assessment of profits, so that an average profit rate or *normal profit* must be assumed for each sector.

Beyond link R it is assumed that there are no more transport costs, so that prices paid are equal to production costs plus profits:

$$^{R}P_j^n = K_j^n + e_j^n \tag{5.39}$$

where $^{R}P_j^n$ represents the price per unit of commodity paid by consumers of sector n in region j in link R. In link $R - 1$, demand units buy from several sectors and regions, and therefore they will pay on average:

$$^{R-1}P_i^m = \frac{\sum_j \sum_n {}^{R-1}X_{ij}^{mn}({}^{R}P_j^n + c_{ij}^n + e_j^n)}{^{R-1}X_i^m} \tag{5.40}$$

In other words, the average unit price for commodity m paid by a purchasing sector n in i will be equal to the quantity purchased multiplied by price components in each case, divided by the total amount of commodities produced in the purchasing sector, $^{R-1}X_i^m$. Similarly, in any link r, the unit price will be:

$$^{r}P_i^m = \frac{\sum_j \sum_n {}^{r}X_{ij}^{mn}({}^{r+1}P_j^n + c_{ij}^n + e_j^n)}{^{r}X_i^m} \tag{5.41}$$

and so on, until the first link is reached, where:

$$^{1}P_i^m = \Sigma_j\Sigma_n {}^{1}X_{ij}^{mn}({}^{2}P_j^n + c_{ij}^n + e_j^n)/Y_i^m \tag{5.42}$$

The multi-region input–output model described above must be solved iteratively. First, an initial estimate of the set of prices must be made. On this basis, distributions from link 1 to R are calculated *forwards* using equations (5.34) and (5.35). On reaching link R (determined through the estimation of convergence), prices are calculated *backwards*, using equations (5.40) to (5.41), starting from link R until link 1 is reached. The process is repeated several times until price convergence is achieved. Note that the model must work with two convergence indicators: one for the generation/distribution of production, and one for the equilibration of prices. This calculation process has been adopted by program REGINP.

Alternatively, the composite cost equation (4.11) can be used to estimate an average perceived set of prices, which in this case becomes:

$$^{r}\tilde{P}_i^{mn} = \frac{1}{-\beta^n}\ln\left[\sum_j \exp\left(-\beta^n\,{}^{r+1}P_j^n + c_{ij}^n + e_j^n\right)\right] \tag{5.43}$$

where $^{r}\tilde{P}_i^{mn}$ is the composite cost perceived by m when purchasing from sectors n to produce a unit produce. Since m requires produce from other sectors as well, total perceived cost for all inputs of sector m in region i will be:

$$^{r}\tilde{P}_i^m = \sum_m a^{mn}\,{}^{r}\tilde{P}_i^{mn} \tag{5.44}$$

This way of calculating prices is not only consistent with random utility theory, but is also easier to calculate. The process begins with link R backwards, because in order to apply equations (5.43) and (5.44) there is no need to know how purchases are going to be distributed. Once link 1 has been reached, calculation turns forward estimating the resulting distributions until the last link is reached, where the process ends without the need for further iterations.

As can be seen, the first approach works on the basis of strict prices that are actually paid in the real system, thus representing real economic transactions. The second approach works on the basis of perceived costs, and hence is subject to interpretation. Such an interpretation is very similar to the one described in chapter 4 when general random utility theory was presented. Whether the criterion of perceived or composite costs is applicable to productive sectors is, however, a matter of debate. The two approaches can be combined: composite costs are calculated to estimate the distributions, and strict costs are simultaneously evaluated to match economic accounting.

5.4.3 Price accumulation in the production chain – abnormal profits

In the model described in the preceding section, it was assumed that normal profits were extracted in every link along the production chain.

There are, however, circumstances in which such profits can be altered, generating *abnormal* profits, which can be higher than the normal ones. In particular cases, these abnormal profits can be called *rents*.

The magnitude of normal profits is mainly determined by the social relationships prevailing in the economic system. Such relationships, in turn, are regulated by the political system which will determine which proportion of the value added in each productive process can be appropriated by capital. In the above formulation, e^n represents capital gains, as an average by sector, determined by the socio-political system.

Abnormal profits or rents may emerge, however, under special circumstances. The two most important conditions under which they can emerge can be defined as *scarcity* and *monopoly*, generating scarcity rent and monopoly rent respectively.

Scarcity rent will emerge when the supply of a particular good in one or more regions cannot grow at the same rate as demand for it. Typical examples are several raw materials and non-renewable natural resources. In an urban area, the most typical example is land. Often, it might be a case of short-term supply restriction, which can be equilibrated in the long term through investment. Labour, for instance, can be considered as a short-term restriction in some fast growing regions, to be equilibrated in the long run through migration.

Monopoly rents arise when a single economic entity has full control of supply and determines the amount to be produced. If it is a private monopoly, the magnitude of production will be determined such that profits can be maximised. The latter, in turn, is a function of the scale of production (determining costs) and the elasticity of demand (determining the number of units to be consumed). These two variables will determine an optimal point where profits are maximum. If the monopoly is public, it is understood that prices are determined such as to maximise total social benefits.

Thus, scarcity and monopoly rent can be seen to be equivalent from an analytical point of view. If supply is fixed in the short run, prices can go above or below normal gains. In general terms, the model proposed in the preceding section should include the assumption that, eventually, any sector and region could become scarce. Two main conclusions can be derived at this point:

- Abnormal rents can be different from one region to another depending on the particular relationships between supply–demand in each region and transport costs.
- Abnormal rents can be different from one time period to another.

Each one of these issues will be dealt with in turn. To begin with, a new variable must be introduced to the system: Q_j^n, the installed capacity of production of each sector n in each region j. Next, the calculation sequence must be modified as follows:

(a) Calculate composite costs along the chain backwards, using equations (5.43) and (5.44), assuming normal profits, denoted by e_j^n.

(b) Assign purchases from demand sectors and regions to supply ones using equations (5.34) and (5.35).
(c) Compare total production assigned to each sector and region with the given Q_j^ns; if assigned production is bigger than production capacity, increase profits e_j^n, reduce otherwise.
(d) Go back to step (a) to recalculate composite costs, thus proceeding iteratively until production assigned is equal to the given Q_j^ns for all regions and sectors.

This solves the problem of abnormal gains in the short run, where installed capacity Q_j^n is assumed to be fixed. The model will result in different profit rates in each sector and region, a fact that can explain many of the existing regional differences, as a combination of transport costs and investment.

In the long run, the values of Q_j^n should be allowed to vary also. To represent this, a dynamic structure must be introduced, for which the following criteria are proposed. Firstly, time must be divided into discrete intervals denoted as t, $t + 1$, $t + 2$, and so on. Secondly, assume that money available for investment is not tagged to any particular sector or region; producers can freely move this money where they believe it will render the highest profits. Thirdly, producers must anticipate where profits will be higher in a following time period, say $t + 1$, but the only piece of information they have in order to assess this is the level of profits in the current time period t, ${}^te_j^n$. If ${}^{t,t+1}M$ is the total amount of investment available between t and $t + 1$, a typical distribution will be:

$$ {}^{t+1}M_j^n = {}^{t,t+1}M \frac{{}^te_j^n}{\sum_{n,j} {}^te_j^n} \tag{5.45} $$

In order to transform this into installed capacity, a simple relationship can be established:

$$ {}^{t+1}Q_j^n = {}^{t+1}M_j^n s^n + {}^tQ_j^n \tag{5.46} $$

where s^n represents the increase of installed capacity in sector n that can be achieved with a unit increase in investment.

5.4.4 *Elasticity of the demand and substitutions in the production chain*

In the preceding sections the problem of prices was discussed. Prices, however, affect the amount of goods that is consumed through elasticity. Assume that the first stage in the calculation process has been performed, estimating the accumulated prices backwards along the production chain. Once the first link in the chain has been reached, i.e. final demand, the number of goods and services to be consumed can be calculated as a function of the accumulated prices, instead of using the fixed a^{mn} coefficients. If prices are high, a small amount will be consumed, and vice versa, which implies a decaying demand function with a slope depending on the nature of the goods and the particular sector of the final demand. In the case of private or public consumption, people or the government will consume smaller amounts of a commodity if the price of the commodity increases. In the case of exports, as prices increase, the country will become less competitive in the world market, and hence will be able to sell fewer goods.

Note that it is implied in this proposition that elasticity of the demand only occurs in the first link (final demand) of the chain, and not in intermediate purchases. The basis for this argument is that the amount of a particular commodity to be consumed is a decision made by consumers, internal or external, and not by the intermediate sectors. The latter can only try to minimise costs, but cannot decide how much will be purchased.

Formally, it was stated that demand for a particular good n was obtained by multiplying final demand by the technical coefficients which were assumed to be fixed. The difference is that now such quantities will be obtained by $Y_i^m \tilde{a}^{mn}$, where $\tilde{a}^{mn}$ is a set of variable technical coefficients applicable only for the first link in the chain, calculated as:

$$\tilde{a}^{mn} = f({}^1P_i^{mn}, \alpha^{mn}) \tag{5.47}$$

where ${}^1P_i^{mn}$ is the accumulated prices in the first link as calculated in equation (5.42), and α^{mn} is the elasticity of the demand of sector m for goods produced by sector n. Alternatively, ${}^1P_i^{mn}$, the composite costs as calculated in (5.43), can be used.

The final form of the calculation sequence when elasticities are introduced is as follows:

(a) Calculate composite costs along the chain backwards, assuming normal profits.
(b) Calculate final demand using variable $\tilde{a}^{mn}$ from equation (5.47).
(c) Assign purchases from demand sectors and regions to supply ones using fixed a^{mn} coefficients.
(d) Compare total production assigned to each sector and region to given capacities; modify profits accordingly.
(e) Go back to step (a) to recalculate composite costs, thus proceeding iteratively until production assigned is equal to the given capacities for all regions and sectors.

Note that final consumption will be different now from one region to another, and that all flows between regions and sectors can be modified as a result of the introduction of elasticities. Program REGINP can help to visualise the effects of elasticities.

The last element to be included in this analysis is that of substitutions. Due to its complexity, however, it will not be included in the formal proposition. Substitutions can be defined as the change in the quantity demanded of a good which results purely from a change in its price relative to the prices of other goods. In contrast to the phenomenon of elasticity, substitutions can occur at any link along the production chain. There is, however, an important difference between substitutions in the final demand with respect to intermediate ones. Substitutions in the final demand can occur between any two goods; for instance, consumers can decide to buy fewer books about land use and transportation and more shoes, if the price of books increases relative to that of shoes, even if books and shoes do not perform similar functions. In the case of intermediate demand this phenomenon can occur only between *substitutes*, that is, between goods that do perform similar functions; for instance, if the price of ivory increases, a manufacturer of pianos can replace ivory with plastic in the manufacture of keys in order to reduce costs, but he will not increase the consumption of other materials from which keys cannot be manufactured. These can also be called *close* substitutes. Another good example of intermediate substitutions is the increase in the purchase of machinery if the price of labour increases.

5.5 Conclusions

In this chapter, macro-economic models have been reviewed; in particular, the input–output model as the most general and useful representation of a spatially interacting economic system. Capitalising on random utility principles, a particular form of a multi-region model has been proposed, which can be useful for the representation of complex urban or regional systems. The important concepts of price accumulation, elasticities and substitutions were introduced to the proposed model. The concepts of normal and abnormal prices were discussed as a result of the introduction of short-term restrictions in the capacity of production in particular sectors or regions. This, in turn, gave rise to a particular dynamic representation of the system.

The resulting general model can be used in many ways at different scales, such as urban or regional. In the following chapter, the general model is applied to the case of the location of activities in space. The transportation component of the system, which plays a role of paramount importance, affecting all aspects of the socio-economic system, requires special treatment, which is left to a final chapter.

6

Urban and regional activity allocation

The previous chapter dealt with a general spatial accounting model which can be applied to many systems. In this chapter the particular application to the location and interaction of activities in the urban or regional space is discussed. It begins by describing the general dynamic structure of the activity system, identifying three main elements: activity allocation and land market, the dynamic relationship between activities and floorspace, and the dynamic relationship between activities and transport. Because of its very particular characteristics, the latter is described in a following chapter. Reference is also made to section 1.5 where general model dynamics were discussed.

6.1 General urban–regional dynamics

In the previous chapter, when discussing the issue of abnormal prices, a particular way of representing supply–demand dynamics was proposed when there are short-term restrictions in the capacity of production in particular sectors and regions. If demand for such goods in certain places exceeds capacity, short-term prices rise, generating abnormal profits or rent. Different levels of profit in particular sectors and regions constitute the determining factor in the distribution of investment in a following time period; given a total amount of money available for investment, an investment allocation model can be constructed to distribute money to sectors and regions. This, in turn, will determine increments in the capacity of production for the next time period, thus generating new demand/supply relationships.

In the case of the location and interaction of activities in space, the system can be viewed as people performing activities in particular places in what can be termed the *activity system*. In such a system, whether at the urban or regional scale, the main restrictions that generate abnormal costs are labour, land/floorspace and transport. It will be assumed, then, that the supply of these three elements is limited in the short run, and will vary in the long run, through investment in the case of land/floorspace and transport, and migration in the case of labour. These demand/supply relationships generate a dynamic structure, as shown in figure 6.1.

The activity system represents a demand for these three elements, but

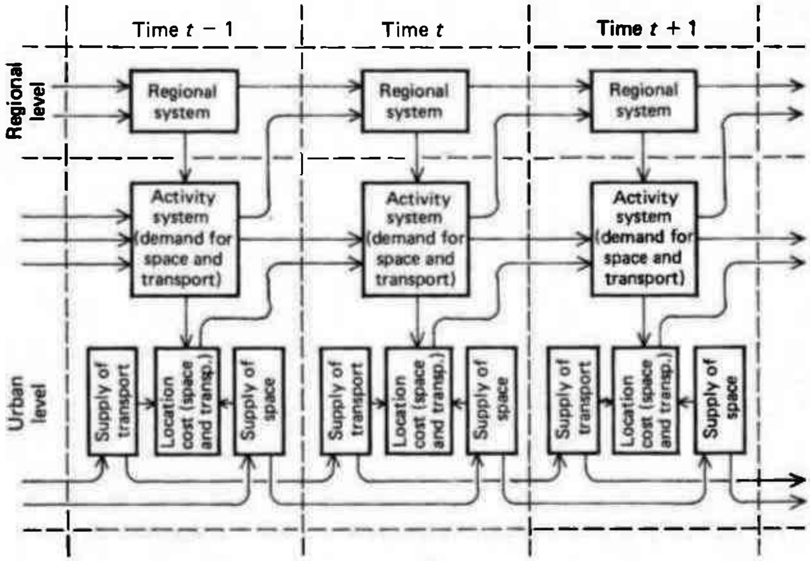


Figure 6.1. Dynamic relationships in the activity system

the characteristics of each one vary. The most typical is land/floorspace, seen as a combined commodity. In a regional system, for instance, cities and their rural hinterlands represent the current supply of land and floorspace, which will be considered as fixed in a particular period of time. The activity system will demand land and floorspace in particular places, and the demand/supply relationship will determine abnormal property prices or land values. In a following time period, investors introduce new supply, more of it where it proved to be more profitable in the past, thus affecting prices.

The above description fits the supply of floorspace quite closely in real terms, although the public sector can distort the system with extra supply where it believes it is necessary. The supply of land, however, is more subject to state intervention, because of regulations and because the supply of land also implies the supply of services, such as sewage, water, electricity and other public utilities.

The supply of transport is made up of basic infrastructure, mainly roads, and transportation systems or operators, such as buses, cars, planes and ships. Since the magnitude of the required investment is so large, the construction of basic infrastructure is generally carried out by the state.

The supply of labour is different in nature in that it is a decision made by large numbers of individuals willing to move from one place to another looking for improved living conditions. In a city, the state is unable to affect changes in the supply of labour directly, and the same applies to regions within a country. The state will, however, exert powerful controls at an international level, regulating the number of people that are allowed to enter a country.

6.2 The activity system

Following the conclusions of chapter 5, the activity system must be represented as a spatial accounting or input–output model. There have been a number of attempts to apply input–output formulations to activity systems, particularly to the urban scale. A very good starting point was provided by Broadbent (1973), although his proposition is based on activity analysis, rather than input–output. Activity analysis has a structure very similar to input–output, but the central matrix is arranged in a different way: rows represent commodities and columns represent activities, the last column representing constraints. The activity–commodity arrangement simplifies the application of linear programming methods, with which it is often confused. Activity analysis is, however, a model, while linear programming is a technique or method for solving numerical systems.

In activity analysis, commodities are the entities of the system, and are produced or consumed by the various activities. Activities, in turn, are transformations of commodity inputs into commodity outputs. The relationship between activities and commodities are represented in a commodity-by-activity matrix a^{mn}, representing the amount of input or output of commodity m required or produced by activity n. A positive entry represents an output, a negative entry represents an input. Constraints are given exogenous elements, and may include basic inputs (negative) or final demands (positive).

The crucial point in Broadbent's proposition is that spatial interaction models can be identified with the coefficients in activity analysis. In doing this, it becomes possible to apply activity analysis to urban areas, and furthermore to use the dual for evaluation purposes. Thus, the author concludes, the direct model can be used to determine the distribution of activities to meet given demands, and the dual model to determine the distribution of costs (or benefits). The use of the dual in evaluation is not pursued here or in Broadbent's proposition, but the form of the resulting indicators gets very close to the random utility theory of composite costs.

There have also been attempts to represent the Lowry model in an input–output form. An important antecedent to this line of research is a paper by Garin (1966), in which the Lowry model is expressed in matrix notation, bringing it very close to an input–output formulation. The same was pursued by McGill (1977), who took Broadbent's proposition as a starting point and interpreted the Lowry model as an input–output model. McGill's proposition is not only a reinterpretation of Lowry's original model, but also an extension to account for all possible interactions between activities. A similar approach is adopted in the following sections.

With the elements provided in the previous chapter, the definition of a general activity allocation model is a relatively straightforward matter. The main difference between an activity system and a general input–output formulation is that each sector in the economy is defined in terms of the number of people involved in the performance of production, i.e. jobs. The definition can be extended further to include all kinds of activities, productive or not, such as residents. Employment and population thus become the main variables.

As in the input–output model, a distinction must be made between final demand sectors and intermediate ones, which can also be termed as basic and induced activities. These activities give rise to a number of functional relationships, similar to the flows of commodities in the original input–output formulation. In this case, though, a typical flow represents the number of activities in a particular sector and zone generated by an origin activity and zone. The rest of the system remains identical to the general random utility based model proposed in section 5.4.1, with identical iterative solution.

For convenience, equation (5.34) is repeated here:

$$ {}^{r}X_{ij}^{mn} = {}^{r-1}X_i^m a^{mn} \frac{\exp(-\beta^n V_{ij}^n)}{\Sigma_j \exp(-\beta^n V_{ij}^n)} \tag{5.34}$$

where ${}^{r}X_{ij}^{mn}$ now represents the number of activities of sector n in zone j generated by activities m in zone i, and where r denotes the iteration number. The utility function can adopt a variety of forms, but for an activity allocation model, the following form is proposed:

$$ V_{ij}^n = c_{ij}^n + \tau^n r_j \tag{6.1}$$

where c_{ij}^n is the composite cost of transport for activity n from i to j, r_j is the value of land in j and τ^n is a parameter regulating the effect of land values on the location of activity n, which combines with β^n, for the effect of composite cost of transport. It is implied that land values exert a negative attraction on activities because they impose a higher cost to households or firms.

The overall structure of the activity model can be seen in figure 6.2, where reference is made to previous and current time periods. Given inputs to the model referring to the previous time period are the location of activities, land values, and composite transport costs. Current time period inputs are the amount of land in each zone, and the net growth of basic activities and of floorspace between the previous time period and the current one.

With this information, the first stage in the calculation of the activity model is the allocation of the given increments of basic activities and of floorspace to zones. These increments, which can be positive or negative, must be added to the previous time values, to obtain current totals which will be considered as fixed supply. The location of the increments of basic employment can be estimated with a simple probabilistic model, but the distribution of the increments in floorspace must include particular variables which will be described in a later section.

Once the number of basic activities and the amount of floorspace in each zone have been determined, the main activity allocation process can be initiated through successive applications of equation (5.34), together with utility functions (6.1), to determine the location of

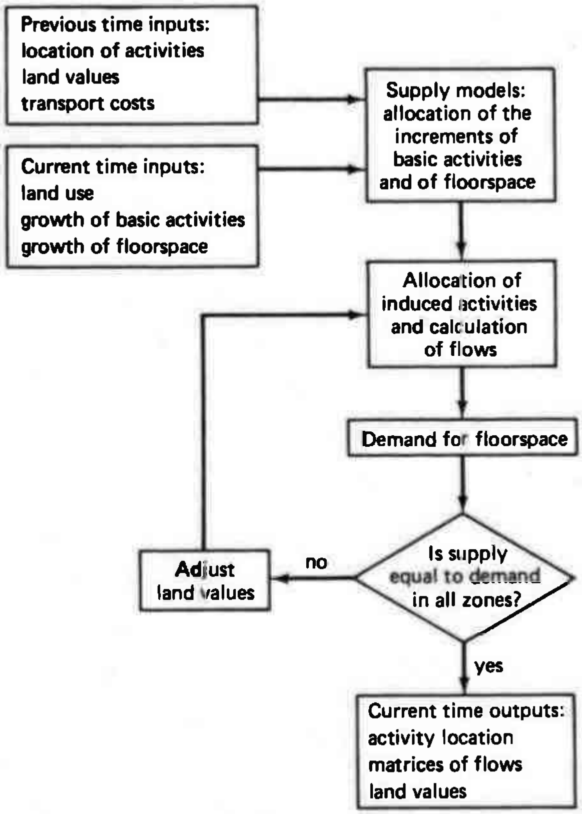


Figure 6.2. General structure of the activity system model

induced activities. Composite transportation costs refer to the previous time period, and their calculation will be explained in some detail in the following chapter. Equation (5.34) will also produce estimates of the flows between activity pairs.

Once all activities have been allocated to zones, total demand for floorspace can be estimated for each zone with appropriate floorspace demand functions. Floorspace demand, $\hat{F}_j$, can now be compared to the current fixed supply of floorspace, F_j. If demand is higher than supply, land values must increase; decrease otherwise. The following adjustment can be performed:

$$ {}^{\sigma+1}r_j = {}^{\sigma}r_j \frac{\hat{F}_j}{F_j} \tag{6.2} $$

where σ denotes the iteration number. With the new land values the activity allocation process starts a fresh iteration. The process is repeated over and over, until $\hat{F}_j$ approximates F_j for all zones.

The main outputs of the activity model are the location of activities in each zone, the supply of floorspace (approximately equal to demand), land values, and a set of X_{ij}^{mn} matrices, the result of applying equation (5.34) in the last iteration, representing the functional relationships between sectors and regions. In the next chapter, these flows will be used as the basis for the calculation of the demand for transport.

The activity model described above can be applied to urban or regional scales alike. In an urban application, the role of land values is quite clear: parts of the city that become very attractive because of accessibility and other considerations end with high land values, and, as a result, acquire more activities with a low sensitivity to land values (low τ'') pushing high sensitivity activities towards cheaper zones. The outcome of this process is change in land use, with activities such as commerce and low income population consuming small amounts of floorspace on expensive land. High sensitivity activities, such as high income population, locate in cheaper zones, consuming larger amounts of land and floorspace.

On a regional scale, land values, which in this case would represent city-wide averages, would be high in attractive regions, so that in the long run certain types of activities would tend to move out. This phenomenon would affect some types of heavy industry and would also affect migration, because even if salaries might be higher, so would the cost of living.

6.3 The dynamic relationship between floorspace supply and the activity system

Two aspects of floorspace are dealt with in this section: demand and supply. The purpose of a floorspace demand model is to estimate the amount of floorspace activities are willing to consume as a function of price. In each time period a certain amount of new-built stock is generally added to the market, and the purpose of a floorspace supply model is to estimate the location of such increments.

6.3.1 Floorspace demand

The relationship between floorspace consumption and land values is determined by a demand function, such as the one shown in figure 6.3,

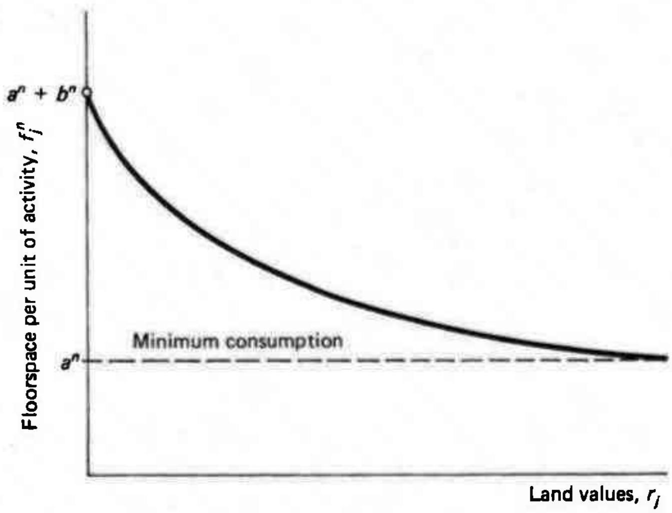


Figure 6.3. Floorspace demand function

where the amount of floorspace a unit of activity n is willing to consume in zone j, f_j^n, is an inverse function of land values. There is a wide choice of functions to represent the elasticity of demand, but a convenient one is:

$$f_j^n = a^n + b^n \exp(-\delta^n r_j) \tag{6.3}$$

where a^n, b^n and δ^n are the parameters of the demand function. Parameter a^n represents the minimum amount of floorspace a unit of activity n is willing to consume in zone j if the land value is ∞; $a^n + b^n$ is the maximum; δ^n is the slope of the demand curve. Total demand for floorspace in a zone will then be:

$$\hat{F}_j = \sum_n X_j^n f_j^n \tag{6.4}$$

where $\hat{F}_j$ is the total demand for floorspace in zone j, and X_j^n is the number of activities of sector n located in j.

6.3.2 Floorspace supply

In order to perform its functions, the activity system requires land and floorspace, two inseparable commodities. However, these two commodities are different in nature. Land is a commodity subject to market transactions like any other commodity, but its supply is strongly determined by state intervention because of the supply of public utilities such as water, sewage, electricity, etc., and because of regulations. Furthermore, a distinction has to be made between potential land available for development, and actual land as market supply. In order for land to become available, public utilities must be provided, together with accessibility. Since these factors depend largely on planning and the construction of infrastructure by the state, land availability will be considered as an exogenous variable to this model.

The actual supply of sites where buildings can be erected can be conceived as a supply factor to be provided through investment. Once infrastructure has been built or brought in, potential land must be developed: plots must be conditioned, earth movements must be made, minor roads must be built, together with all kinds of minor supplies. For analytical purposes these activities will be considered as part of the supply of floorspace. In other words, it will be assumed that the supply of potential land is an exogenous variable subject to land use policy, and the actual provision of sites and buildings will be considered as a commodity subject to market mechanisms under the assumption of restricted supply. The purpose of the paragraphs that follow is to construct a model that can simulate the behaviour of the suppliers of built stock, representing the way they respond to demand, given an existing stock, land availability, and accessibility patterns.

There are two main issues involved in the estimation of the supply of built stock. The first refers to the overall supply of stock in the study area, whether a country at a regional scale, or a region at an urban scale. This subject has been treated by many authors, particularly urban economists such as Muth (1968, 1969), Kirwan and Martin (1970), Goodall (1972) and Mills (1969). The overall growth of the building sector is certainly a complicated issue, dependent on elements such as national economic growth, development of the building industry, government incentives, availability and prices of building materials, financial and credit facilities, income of the population, and so on. For these reasons, the overall level of activity of the building sector will be considered as given.

The second issue involves the spatial distribution of a given increment of floorspace. One of the first models of the distribution of a given increment of floorspace is due to Hansen (1959). Later Echenique, Crowther and Lindsay (1969) proposed a similar model and incorporated it into the Lowry structure to restrict residential location. These early models are taken as a starting point.

What is required is the definition of a utility function, assuming a rational behaviour of the suppliers of built stock. The net increment in the supply of built stock at a particular time period can then be allocated to zones according to an a-priori probability and to the relative expected benefit that suppliers perceive in each location.

The a-priori probability can be defined as the *potential floorspace* that can be built in each zone, i.e. the difference between what is allowed by local regulations in each zone, and the amount of floorspace already located:

$$ {}^{t+1}H_j^P = {}^{t+1}H_j^R - {}^{t}H_j \tag{6.5} $$

where ${}^{t+1}H_j^P$ is the potential new floorspace for the next time period in zone j, ${}^{t+1}H_j^R$ is the maximum amount of floorspace that will be allowed by regulations in each zone and ${}^{t}H_j$ is the floorspace already existing in each zone. Equation (6.5) assumes that developers know in advance the building regulations for the next time period.

Utility to producers can be estimated as the difference between selling price r_j in each zone, and production costs, b_j. Then:

$$ {}^{t+1}U_j = {}^{t}r_j - {}^{t}b_j \tag{6.6} $$

where ${}^{t+1}U_j$ is the expected utility to producers in zone j at time $t + 1$. The production cost ${}^{t}b_j$ must be calculated outside the model, to include elements such as the actual cost of building, infrastructure, taxation, and so on. Density affects the cost of building, since high rise blocks are more expensive to build than low density houses. Note that both the selling price and the cost refer to the current time period; this is because suppliers cannot know future prices for sure, but they do know the current difference between prices and costs, the only basis for their planning. New built stock in each zone can then be calculated as follows:

$$ {}^{t+1}\mathrm{d}H_j = \frac{{}^{t+1}\mathrm{d}H \; {}^{t+1}H_j^P \; {}^{t+1}U_j}{\Sigma_j \, {}^{t+1}H_j^P \; {}^{t+1}U_j} \tag{6.7} $$

where ${}^{t+1}\mathrm{d}H_j$ is the amount of new built stock to be supplied in zone j

in the next time period, and where ${}^{t+1}\mathrm{d}H$ is the given system-wide increment in floorspace planned for the next time period. Total future floorspace in each zone will be:

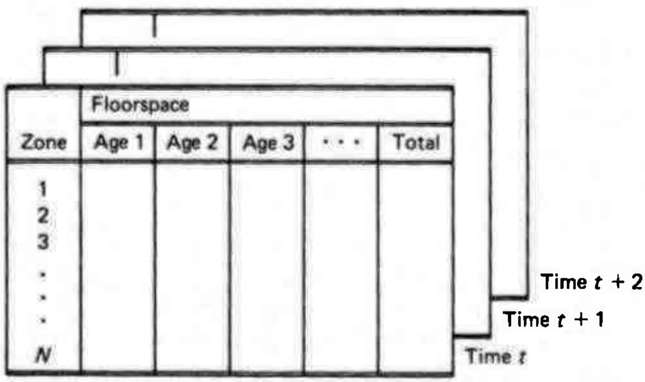
$$ {}^{t+1}H_j = {}^{t}H_j + {}^{t+1}\mathrm{d}H_j \tag{6.8} $$

The model that has been described includes a number of simplifying assumptions, the main one being the consideration of only one type of floorspace of the same age. Floorspace can be divided into different types, such as residential, commercial, industrial, and so on. In turn, residential floorspace can be disaggregated into, say, detached, terrace and flats, or even by size and quality. Anas (1982) has looked into this issue in detail, introducing the concept of *sub-markets*. From the point of view of the supply model, it is simply a matter of disaggregating stock from ${}^{t}H_j$ to ${}^{t}H_j^f$, where f denotes building type. Equation (6.7) becomes:

$$ {}^{t+1}\mathrm{d}H_j^f = \frac{{}^{t+1}\mathrm{d}H\; {}^{t+1}H_j^{Pf}\; {}^{t+1}U_j^f}{\Sigma_j \Sigma_f\, {}^{t+1}H_j^{Pf}\; {}^{t+1}U_j^f} \tag{6.9} $$

Equation (6.9) implies that the total increment in floorspace, ${}^{t+1}\mathrm{d}H$, will be allocated simultaneously to zones and building types according to expected profits in each case, ${}^{t+1}U_j^f$. Note that potential floorspace must be disaggregated by type f also, because regulations can restrict the building of certain types in particular zones.

The age of buildings can also be brought into the analysis. In order to deal with this problem, floorspace can be disaggregated by age group, q. Built stock in each zone and age group, ${}^{t}H_j^q$ can then be arranged in a matrix form of the following type.



This matrix can be operated with a dynamic model based on a Markov chain, to simulate the amount of floorspace in each zone and age group in each time period. The following relationships can be established:

$$ {}^{t+1}H_j^q = {}^{t}H_j^q p^q + {}^{t+1}\mathrm{d}H_j, \quad \text{for } q = 1 \tag{6.10} $$

and

$$ {}^{t+1}H_j^q = {}^{t}H_j^q p^q + {}^{t}H_j^{q-1}(1 - p^{q-1}), \quad \text{for } q > 1 \tag{6.11} $$

Equation (6.10) estimates the amount of floorspace in the first age group ($q = 1$) as the amount of floorspace already existing in age group $q = 1$ in time period t, ${}^{t}H_j^q$, multiplied by the probability that the built stock will stay in the same age group, p^q. To this amount of 'surviving' stock in the first age group, the total amount of new built stock, ${}^{t+1}\mathrm{d}H_j$ must be added, since the new stock is naturally 'born' to the first age group.

Equation (6.11) states that, for all age groups other than the first one, the amount of floorspace in a zone in time period $t + 1$ is equal to the stock that 'survived' in the same age group, ${}^{t}H_j^q p^q$, plus the stock that 'aged' from the previous group during the period, ${}^{t}H_j^{q-1}(1 - p^{q-1})$. Note that p^q must be less than 1. Note also that in every time period there is a certain amount of stock that disappears altogether, at a rate of ${}^{t}H_j^Q(1 - p^Q)$, where Q is the oldest age group.

A further improvement can be introduced if demolitions are taken into consideration. Figure 6.4 makes a distinction between different types of land: a certain amount of land is currently occupied by buildings, and the rest remains as available vacant land. The shaded area represents the existing floorspace, ${}^{t}H_j$. The new stock to be added for the next time period can be of two types: new stock built on vacant land ($v = 1$) and new stock built on previously occupied land ($v = 2$). The cost of building will be different in each case, because the cost of demolitions will have to be added to the $v = 2$ type. In this way, new built stock becomes ${}^{t+1}\mathrm{d}H_j^v$, where v denotes the vacancy condition. Equation (6.9) now becomes:

$$ {}^{t+1}\mathrm{d}H_j^v = \frac{{}^{t+1}\mathrm{d}H \; {}^{t+1}H_j^{PJ} \; {}^{t+1}U_j^v}{\Sigma_j \Sigma_v \, {}^{t+1}H_j^{PJ} \; {}^{t+1}U_j^v} \tag{6.12} $$

Equation (6.12) represents a choice for the suppliers of built stock: build on vacant land or replace old buildings by new ones. There will be a different cost and hence a different utility associated with each choice, ${}^{t+1}U_j^v$, and different a-priori probabilities, ${}^{t+1}H_j^{PJ}$, represented by the potential floorspace in each case. The selling price in each case will be the same, since potential buyers will be indifferent with respect to

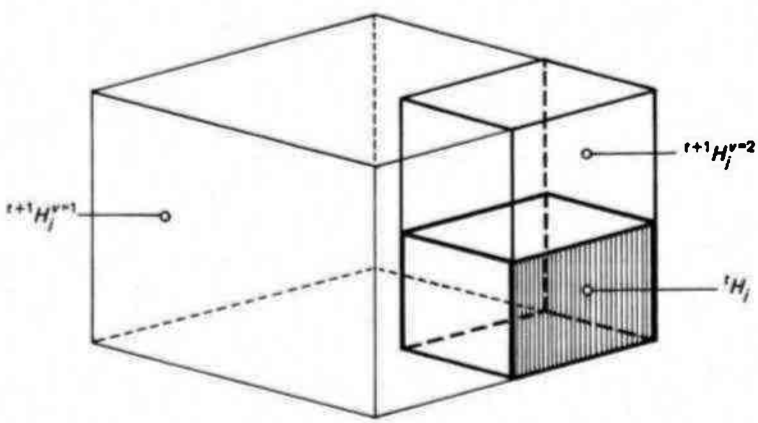


Figure 6.4. Different types of potential floorspace

which vacancy condition existed before the erection of the present buildings.

In order to simplify the presentation, a number of issues have been left out. In particular, densities have remained implicit and have been mentioned only as influencing the cost of buildings and demolitions. Also, the concept of sub-markets can be elaborated along the lines developed by Anas (1982), whereby some activities must be restricted to certain types of buildings, such as industry, commerce, schools, and so on. Public housing can also be considered in the model, to be added to total stock, with a residential restriction.

Another important issue that must be included, particularly in developing countries, is the inclusion of an informal housing sector. In many cities of the third world, the informal sector constitutes a large proportion of the total stock. A possible way to include this begins by modifying the floorspace demand function (6.3) by removing the minimum consumption parameter, a^n:

$$f_j^n = b^n \exp(-\delta^n r_j) \tag{6.13}$$

Equation (6.13) still represents the amount of floorspace a unit of activity n is willing to consume given that r_j is the price of land. In this case, however, the demand function converges to zero as r_j increases. Since activities cannot consume below the minimum a^n, the model must then proceed as follows. In the first iteration all population is assumed to be within the formal sector. Once all activities have been allocated to zones, demand functions (6.13) must be computed. Then all population whose demand for floorspace goes below the minimum a^n are assigned to the informal sector and allocated to zones with a separate sub-model. In a second iteration, since there is a smaller population competing for the same amount of formal floorspace, prices will drop, and consequently part of the informal sector can be brought back into the formal sector. This process continues until a state of equilibrium is reached.

In the dynamic structure, if floorspace grows at a smaller rate than population growth, the informal sector will grow. Public housing will increase supply, and hence reduce prices, thus reducing the proportion of informal population. At a regional level, this model can explain the different proportions of informal population between cities.

6.4 Migration between regions

In a country, people move from one region to another for a number of reasons, the main one being the search for better employment opportunities. There are, however, many other reasons, such as the search for better services, a more favourable cost of living, weather and environmental considerations, cheaper housing, and so on. In an urban area, people move from one location to another for similar reasons, but the weights given to each factor are different. An important factor restricting changes in the location of population is the cost of moving, both economical and psychological. This friction will affect more regional movements than urban movements, because both types of costs are higher in the former case.

In traditional demographic analysis, migration movements at a regional scale are represented through in or out migration rates. Such is the case in the so-called cohort survival model, well documented in Rogers (1968). A cohort is simply a category of population from one or several points of view, mainly age and sex. From known data, the rate at which population migrates in or out is calculated for each cohort, and when age is considered, migration rates are combined with survival probabilities. Based on historical records, these rates are projected into the future. The result of such a model will be the number of people that move in or out of each region.

This approach can be criticised from many points of view. A migration matrix cannot be calculated, and some fundamental variables, such as job availability and accessibility between regions, are not recognised. For this reason, spatial interaction models became popular quite early on, with the work of Zipf (1949), Lowry (1968), Masser (1969) and Wilson (1974). Wilson, in particular, integrated this approach with the matrix methods developed by Rogers (1968) to represent age and cohort survival probabilities.

In the activity allocation dynamic model described in section 6.2 above, migration is represented implicitly, whether applied to a regional system or to an urban area. Population is allocated to zones or regions in each time period as a function of the location of employment, services, accessibility patterns and land values. Migration is then, simply the difference in residential location between one time period and the next.

It can be argued, however, that this implies an instant and costless migration. As soon as there is a change in one of the variables affecting residential location, people are assumed to move immediately to new, more convenient locations. There are, though, some built-in limiting factors in the model that restrict the speed of such movements. On the one hand, employment location changes relatively slowly, restricted in turn by factors such as availability of raw materials, markets, and transportation costs from its suppliers. On the other hand, changes in the supply of floorspace are even slower. In the floorspace location models (6.9) or (6.12), an incremental approach was adopted; hence, a large proportion of the stock remains unchanged from one time period to the next. As a result, land values in particular zones may increase, but it will take some time before the cost of housing forces population to change location. In the activity model of equation (5.34) and (6.1) the relative importance of land values as against accessibility to employment in determining changes in residential location is regulated by the relative magnitude of parameters β^n and τ^n. Further inertia can be added to the activity model by including an attractor variable W_j^n to equation 5.34), thus becoming:

$$ {}^{t}X_{ij}^{mn} = {}^{t-1}X_i^m a^{mn} \frac{W_j^n \exp(-\beta^n V_{ij}^n)}{\Sigma_j W_j^n \exp(-\beta^n V_{ij}^n)} \qquad (6.14) $$

where W_j^n can be defined to represent previous time population; since population can be of different socio-economic characteristics, W_j^n can be further refined by including the weighted sum of each group in the previous time period.

Even if the above approach is a convenient way of representing migration, it does not provide an explicit migration matrix. This, however, can be obtained by adding a post-simulation calculation of the form:

$$M_{ij}^n = {}^tR_n \frac{({}^tR_j^n - {}^{t-1}R_j^n) \exp(-\beta^n c_{ij}^n + c^{*n})}{\Sigma_j({}^tR_j^n - {}^{t-1}R_j^n) \exp(-\beta^n c_{ij}^n + c^{*n})} \tag{6.15}$$

where M_{ij}^n represents migration of group n from i to j, ${}^tR_n$ is total population of type n, $({}^tR_j^n - {}^{t-1}R_j^n)$ represents the difference between the residential location in periods t and $t - 1$, and where c^{*n} represents a fixed cost of migration for group n. M_{ii}^n represents people that do not migrate.

6.5 Spatial hierarchies

In many cases it will be useful to distinguish two or more hierarchical levels when dividing the study area into zones. If it is a regional application, the analyst will probably divide the whole country into regions, but at the same time he might be interested in dividing some of the regions into smaller spatial units. The same will apply to an urban application because the interest of the study centres around certain parts of the city. It could also be that the study covers the whole city, but there are important relations between the city and its surrounding region; in this case, the analyst can define a first hierarchical level representing regions, of which the city is one, and then divide the city in second-level zones in the usual way. In chapter 8 several applications of this kind are described.

In order to represent this in the model of the activity system, equation (6.14) must be modified as follows:

$$\begin{aligned} {}^rX_{ij}^{mn} = {}& {}^{r-1}X_i^m a^{mn} \frac{{}^1W_J^n \exp(-{}^1\beta^n\, {}^1V_{iJ}^n)}{\Sigma_J {}^1W_J^n \exp(-{}^1\beta^n\, {}^1V_{iJ}^n)} \\ & \cdot \frac{{}^2W_j^n \exp(-{}^2\beta^n\, {}^2V_{ij}^n)}{\Sigma_j {}^2W_j^n \exp(-{}^2\beta^n\, {}^2V_{ij}^n)}, \quad \forall j \in J \end{aligned} \tag{6.16}$$

where indices 1 and 2 denote corresponding hierarchical levels, and where J denotes a first-level zone and j a second-level zone. The expression $j \in J$ means that in the second part of equation (6.16) summation is made over all second-level zones j into which the first-level zone J has been divided. The interpretation of equation (6.16) is straightforward. If it was being applied to the location of residents of type n, the first probability would assign residents to a first hierarchy of zones; the probability of locating in a second-level zone is then the probability of locating in the first-level zone to which it belongs, over all first-level zones, multiplied by the probability of locating in the second-level zone over all others that belong to the same first-level zone.

6.6 Conclusions

This chapter has dealt mainly with the activity system, that is, the explanation of how activities locate and interact in space. Most of the theoretical elements had already been dealt with in the previous chapter, so it was mainly a matter of applying a general theoretical framework to a particular instance. The result is a model that can be applied at a regional or an urban level or, if hierarchies are used, to a combination of both.

There are, however, very specific issues related to the location of activities, which were discussed. The most important of these is the demand and supply of floorspace. This is, perhaps, one of the most interesting areas of research in land use modelling, because of its relevance in real social contexts and because of its challenging complexity. In this chapter only the main elements of the subject were covered, leaving out a large number of aspects to be tackled in the future. The case of marginal population, of great importance to third world countries, has been only suggested, including the proposition of a formalised model, but remains largely unexplored. Floorspace in Europe and the USA is the largest energy consuming element of society apart from industry; the careful study of within-space energy consumption is of great relevance for integrated land use and transport modelling (see Owens, 1986; Rickaby and de la Barra, 1987).

7

The transportation system

The relationship between the location of activities and the transport system has been discussed extensively in the literature for many years, but it is only recently that it has been established more formally. This is due perhaps to the historical development of theories related to the use of space and those related to the use of transport, and to the fact that both evolved in relative isolation.

It was pointed out in previous chapters that the majority of land use theories considered the transportation system as having a definite effect on the location of activities; this is a common element in the works of Von Thünen (1826), Christaller (1933), Hansen (1959), Wingo (1961), Alonso (1964) and Lowry (1964), as well as most of the land use research that took place in the late 1960s and early 1970s. All of this work, though, considered accessibility or transport costs in an ambiguous way, and basically as exogenous. Even in the work of Wingo, which could be considered to have pioneered the field because it developed both transport and land use to considerable extent, the two remained in separate compartments, and the transport variables that go into the land use model were restricted to the concept of 'distance to the centre'.

In turn, transportation research has traditionally enclosed the land use aspects into an equally ambiguous concept of 'demand', an area whose sole purpose has been to provide socio-economic inputs for the transportation study. This approach results in a strong imbalance between a detailed description of the transport related variables and the gross oversimplification of the activity system, generally a given exogenous variable. However, concern about the importance of land use has been expressed many times by transportation planners, and in this sense the work of Buchanan (1963) became a landmark. Since then the subject has become standard in most textbooks and there are authors who refer to the subject in a very explicit way (e.g. Blunden, 1971). Even so, the subject is only referenced, and cannot be considered as an integrated theory.

After 1970 came the first generation of combined land use and transport models, particularly with the important SELNEC model (Wilson *et al.*, 1969; SELNEC, 1971, 1972) and the works of Putman (1973, 1975a, b, c), and Echenique *et al.* (1973) (see also de la Barra *et al.*, 1974). The main characteristic of these models is that the socio-economic inputs required by a transport model are provided by a land use model, instead of giving them as exogenous data. In turn, the transport model calculates a generalised cost of transport, which is fed back into the land use model. A review of this generation of models, which can all be fitted (some more forcibly than others) into a common scheme called a *linked land use and transport model*, is the subject of the next section. This is followed by a second section where an *integrated scheme* is put forward to solve some of the problems of the linked scheme. Final sections are devoted to the discussion of particular aspects of the transport system, such as perceived costs, trip generation, modal split and trip assignment.

7.1 The linked land use and transport model

Figure 7.1 describes a possible common structure for a typical linked land use and transport model. It corresponds to the one used in de la Barra *et al.* (1974), but presents common aspects with the SELNEC model and the structure proposed by Echenique *et al.* (1973).

The calculation sequence starts with a regional model which consists of two linked sub-models: a regional employment sub-model and a demographic sub-model, which together perform the calculations of the totals of population and employment per region and sector for the urban area to which the model is being applied.

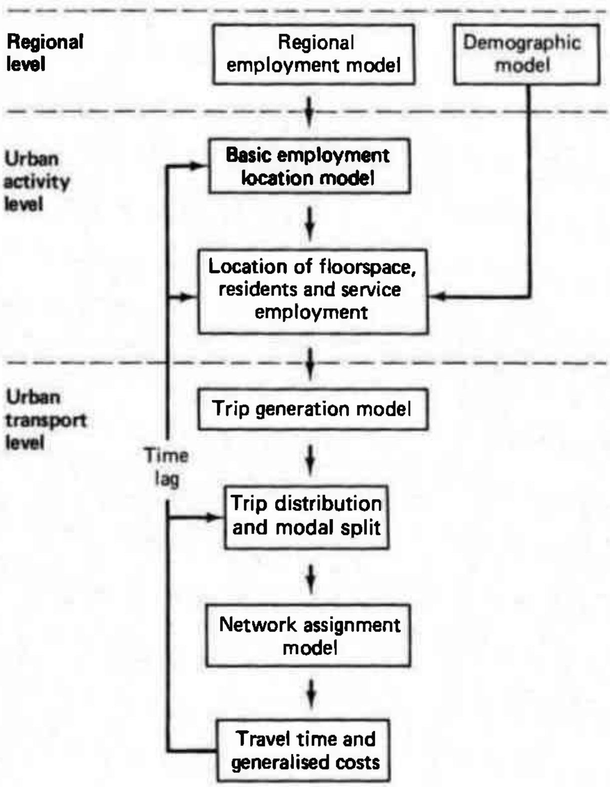


Figure 7.1. General structure of a linked land use and transport model

The next stage corresponds to the location of activities within the urban area, and consists of the location of basic employment, of floor-space, residential population and service employment, from the totals generated by the regional model. These in turn are inputs to the transportation model, consisting of four traditional steps: generation, distribution and modal split, trip assignment, and generalised costs.

The trip generation model transforms the activities by type per zone calculated by the land use model, into trip productions and attractions, that is, the number of trips that originate in each zone and the number of trip ends in each zone. The trip distribution model connects productions to attractions to produce a set of origin–destination matrices. The modal split model separates these trip matrices by mode. Trip distribution and modal split can be combined in a single model, as in equation (3.27). The resulting trips by mode are then assigned to the different routes available in the network by the assignment model. Finally, all these calculations are used to estimate travel times between zones by mode, which are affected by the level of congestion in each link of the network. With travel times and other cost-related parameters, generalised transport costs can be estimated.

From the generalised cost calculations, two main feedbacks are recognised. The first goes back to the trip distribution stage because, as congestion builds up in certain parts of the network, the distribution of trips is affected, and the probabilities of choosing each mode can change. This feedback is equivalent to an equilibrium between supply and demand of transport. Because of the characteristics of the transport system, this equilibrium is assumed to take place instantaneously, that is, no time lag is considered.

The second feedback loop goes back to the location of activities, affected by changes in the generalised cost of travel between zones. This second loop is assumed to take place more slowly, because activities will take some time to adapt to changes in accessibility. An explicit time lag becomes necessary.

As can be seen, the two key elements that relate land use to transportation are trip generation and generalised costs. For the calculation of the number of trip ends as a function of the number of activities in each zone, two methods have been traditionally used: linear regression, and category analysis developed by Wooton and Pick (1967). The linear regression model estimates the number of trip ends as a linear function of the level of activities. In Echenique *et al.* (1973) and in de la Barra *et al.* (1974) trips are divided into car-owners and non-car-owners, and all activity types produced by the land use model are entered into the calculation of the parameters of the linear function. If peak hour trips are being estimated, residential activities generally result with the highest parameters in trip productions, and employment activities for the trip attractions. The set of parameters is estimated through linear regression against real data.

Category analysis methods derive the value of a set of trip generation rates directly by cross-classification. The number of trips made by households are cross-tabulated by different types, including factors like family size, number of working members, car-ownership, income, and so on. Category analysis was developed as a way of getting round certain difficulties in the linear regression model, particularly that of non-linearities. However, as Dale (1977) points out, 'category analysis is a particular form of regression analysis', and the results are very similar. The SELNEC model used a modification of Wooton and Pick's (1967) model, using 108 household categories (income/car-ownership/family structure), and calculated trip attractions from eight land use categories.

In general, the following can be said about these trip generation models:

(a) Trip rates are inelastic with respect to travel cost. If trips are shorter or cheaper, people tend to travel more often, an important factor which is not included in the model. If new transport facilities are introduced, the total number of trips remains the same.
(b) As the number of categories increases, the number of observations from real data tends to decrease sharply, making the results unreliable.
(c) If the general model is to be coupled to the land use model, there is no point in having a large number of categories, because these will not be available in future predictions.

The second land use/transport interface in these models is the calculation of generalised costs of travel. The purpose of this calculation is to provide the land use model with accessibility indicators in matrix form, that is, the c_{ij}^{n} matrices used in previous equations. In the three examples above, these are obtained as simple averages.

7.2 The integrated land use and transport model

The set of models described above is the result of the direct linking of research areas that evolved in relative isolation. Areas like trip generation, distribution, modal split, assignment and so on, were the subject of independent research for many years.

Apart of what has already been said about the trip generation model, there are other deficiencies of the linked approach that can be pointed out. Senior (1977) questioned the validity of the trip distribution model in some detail. The central question regarding this model can be put as follows: if, say, a residential model has simulated the relationship between place of work and place of residence, producing a residence–work flow matrix, why should another similar, though more constrained, distribution of trips to work be performed? In fact, the residential model is based on an origin-constrained model in entropy maximising terms, which is not identical to the doubly-constrained model generally used in trip distribution, though they are performing a similar function. The origin-constrained model assumes jobs as fixed and given, an assumption that is consistent with the overall structure of the model.

In doing so it simulates the X_{ij}^{mn} matrix which would be identical to work–residence trips if the trip rate was equal to one. It is then a relatively simple matter to transform these functional flows into actual trips by using an adequate trip generation model. In these circumstances, the trip distribution model is essentially redundant and ambiguous in its interpretation, and hence must be eliminated from the model structure.

A second aspect of the linked approach that can be criticised is the use of averages for the calculation of generalised costs. This issue was discussed in some detail when describing random utility theory, where it was explained that equation (4.11) was the correct way of estimating composite costs. The difference between simple averaging and the logarithmic average of (4.11) was further demonstrated with a numerical example in section 4.3. It is clear, then, that composite costs must replace averages in the calculations.

However, in section 4.3 a particular way of representing and calculating decision chains was presented. It was argued that, in order to apply composite costs, the calculation sequence should first follow the decision chain backwards, aggregating composite costs by successive applications of equation (4.11), and then reverse the flow forwards to estimate probabilities. Furthermore, in section 4.4, when variable costs and elasticities were introduced, the calculation sequence became iterative.

Such is the case in the land use and transport integration. The process can be viewed as a decision chain, as shown in figure 7.2. At the land use level, activities decide where to locate, as a function, among other variables, of the cost of interaction. In other words, the X_{ij}^{mn} is a function of c_{ij}^{n}, the generalised cost of travel for activity n (note that it refers only to n, the 'moving' activity). X_{ij}^{mn} can be aggregated to X_{ij}^{n} to represent potential travel demand; if this demand is assumed to be elastic, then the actual number of trips, T_{ij}^{n}, is again a function of the generalised cost of travel c_{ij}^{n}.

Modal split, then, can be viewed as a mode choice, thus representing a second link in the decision chain. T_{ij}^{n} is transformed into T_{ij}^{nk}, the

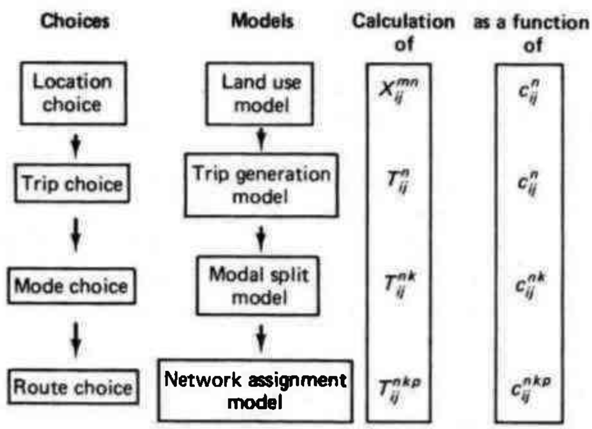


Figure 7.2. Sequence of decisions and sequence of models in the integrated land use and transport model

number of trips by activity n from i to j by mode k, and will be a function of the generalised cost of travel by mode, c_{ij}^{nk}. Finally, T_{ij}^{nk} must be transformed into T_{ij}^{nkp}, the number of trips by activity n from i to j by mode k and route p, as a function of the cost of the corresponding route, c_{ij}^{nkp}. Hence, assignment can be viewed as a problem of route choice.

Following the conclusions of section 4.4, the resulting calculation sequence now becomes:

(a) Start from the last link along the chain, calculating the cost of travel at a route level.
(b) Aggregate costs from route level (c_{ij}^{nkp}) to mode level (c_{ij}^{nk}) and to an origin–destination level (c_{ij}^{n}), which shall be called here the generalised composite cost of travel.
(c) Perform the activity location calculations as a function of generalised composite travel costs.
(d) Estimate the number of trips with an elastic trip generation model.
(e) Calculate modal split probabilities as a function of the composite costs by mode.
(f) Calculate the assignment probabilities as a function of the travel cost at a route level.
(g) Compare demand with supply, in this case, the number of trips assigned to each route with the capacity of the route, and adjust travel times accordingly.

Step (g) is called *capacity restriction*, the process in which travel times are adjusted according to the demand over capacity ratio in each link of the transport network. The resulting increased travel times due to congestion brings the calculation sequence back to step (a), starting a fresh iteration.

The explicit dynamic structure must be kept in mind. In this case, a time lag must be inserted between the location/interaction of activities and the transport chain. Because in the resulting scheme the destination choice has been integrated with the location choice, it is now possible to denominate such a model as integrated rather than linked.

What remains to be discussed are the details of how each of the transport-related calculations can be performed. In the following sections, each step is presented in turn. Because of the strong integration between them, the particular sequence in which they should be discussed is not of great importance. The presentation will begin by covering the issue of network analysis, which lies at the heart of the representation of transport systems. Next, the problem of travel cost is presented, together with that of capacity restriction. This is followed by modal split, and finally route assignment.

7.3 Network analysis for transportation systems

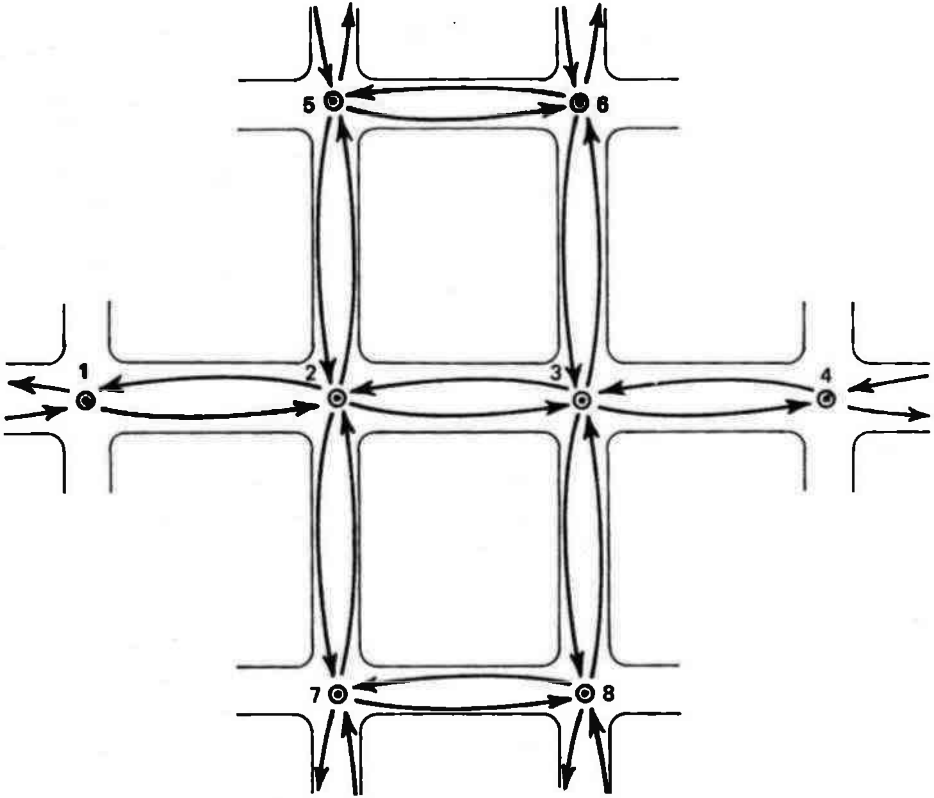
There are many systems that require networks for their representation. Such is the case in electric circuits, water supplies, telephone networks, and so on. Networks are also used to represent the activity sequence in productive processes, an area well covered in operational research. This has resulted in the development of a whole area of mathematics: graph theory. Transportation analysis has drawn heavily upon graph theory, and many of the techniques are direct applications from other areas of research. This, as will be pointed out, can lead to some confusions, because a transport system has many peculiarities, hence the need to develop techniques of its own.

In very general terms, a transport network is defined as a directed graph, that is, a set of nodes connected by directional links, such as the one shown in figure 7.3. In its most simple conception, a transport network represents the physical infrastructure, where links represent roads and nodes represent intersections. This definition can be made more complex with the inclusion of terminal points, such as railway or metro stations, airports, deposits, and so on. Links do not always have to be physical links, since they can represent air or water routes as well. Also transport systems, like a bus system or a metro system, require specific representations, so that links and nodes finally become entities defined by many characteristics.

In an integrated land use and transport model, nodes not only represent critical points in the network where users can travel from one link to another, but they can also represent points where trip ends occur. These special nodes are called *centroids*, and must coincide with the definition of zones in the activity system.

A further complication in the definition of a network is that three main social entities must be distinguished: users, operators and administrators. Users represent demand, and can be people or commodities. Users will *perceive* a cost from each trip they make, which will be a function of a number of variables such as out-of-pocket costs, travel time, parking, and so on. Operators represent individuals or firms that provide a transport service, such as bus companies, freight trains, airlines and so on. Operators charge users a fare and pay operating costs. Administrators are the entities in charge of the maintenance of the infrastructure, a role generally fulfilled by the government or concessionaires. Administrators have to pay for maintenance costs and can charge tolls, or parking fees, or storage charges. There are some peculiar cases, such as the case of automobiles which are operated by the users themselves, and railways where the same institution operates the service and maintains the infrastructure.

There are also some important concepts related to traffic flow that must be taken into consideration in the definition of the transport network. These concepts are volume of flow, density, capacity of a link and speed. Figure 7.4 shows how these concepts relate to each other. Figure 7.4(a) relates volume to density; at zero density, flow will be zero as well as at maximum (jam) density; at some intermediate point, volume will be maximum. The maximum is generally referred to as the capacity of a link. Mean speed is normally defined as flow over density, i.e. the slope of *OA*, *OB*, and *OC*. *OC* is the mean free speed, equivalent to the speed of vehicles when the road is relatively empty. As the density



Network code

Origin intersection	Destination intersection
1	2
2	1
2	3
3	2
2	5
5	2
2	7
7	2
3	4
4	3
3	6
6	3
3	8
8	3
5	6
6	5
7	8
8	7

No. of links: 18
No. of nodes: 8

Figure 7.3. Typical example of a transport network represented as a directed graph

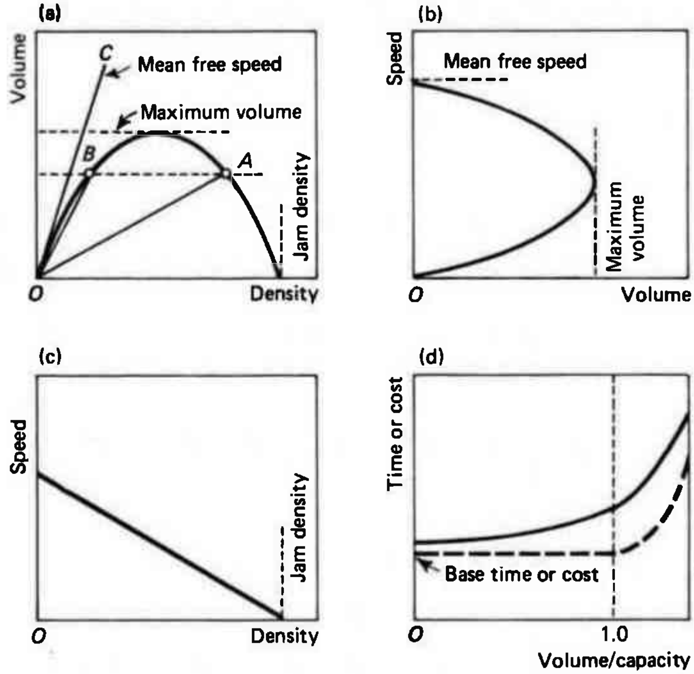


Figure 7.4. Basic diagrams of traffic flow

increases from zero to the maximum, mean speed decreases from the mean free speed to zero.

Figure 7.4(b) is a direct consequence of 7.4(a). If the volume of flow is zero, speed will be maximum, equal to the mean free speed. As speed increases, a larger volume of traffic crosses the link, until a maximum is reached from where speed decreases gradually. Similarly, as the density of traffic increases, speed will decrease also, as in figure 7.4(c), until the jam density is reached, where speed will be zero.

Figure 7.4(d) shows the situation in a particular link of given capacity and mean free speed. As more traffic enters the link, the volume/capacity (V/C) increases until it reaches a maximum of 1.0, that is, when volume equals capacity. The time vehicles take to cross the link increases as the V/C relationship increases, from a minimum determined by the mean free speed. The particular shape of the curve is a matter of empirical testing, and will be discussed in a later section.

As can be seen, the definition of a transport network can be a complicated process involving many issues. Most of the characteristics of the transport network are described in the *network code*, a list of links with their description stored in a computer file. A large variety of formats in which a network can be coded for analytical purposes can be devised, and a convenient one is to include the following items: origin and destination nodes, length of the link, link type, capacity of the link, and capacity of each transport operator that uses the link. The link type code, in turn, can refer to a number of generic characteristics, such as mean free speed by operator, administrator, maintenance costs, operating costs, possible tolls or other charges, and so on.

7.3.1 *Path building*

Once the transport network has been properly represented in the code, one of the first processes that must be carried out is the connection of origin to destination centroids, through a path-building algorithm. Path-building algorithms are strongly related to assignment algorithms, but this issue will be covered in a later section. There are two possible types of such algorithms: single and multiple path search. In transportation analysis the single-path algorithm is always related to the problem of finding the path with the least cost to connect an origin to a destination centroid, while the multi-path search looks at the first *n*-paths.

The minimum path problem can be stated as follows. Given a transport network, find the route with the least transport cost from a given origin to a given destination node. This problem has been very well covered in the literature. Most algorithms are based on a method first developed by Dijkstra (1959). Moore (1963) adapted the method to the transportation problem. The method proceeds in the following stages:

(a) Label all nodes with a very large cost.
(b) Label the destination node with a cost of zero.
(c) Examine all nodes connected to the destination node (candidates) and set cost labels to the actual travel cost from each node to the destination node.
(d) Examine, in turn, nodes connected to the candidates, labelling them with accumulated costs; if it is found that a node has already been labelled, choose the smallest and re-label.

This process is continued until all nodes have been labelled. Computationally, the method is very attractive, because it ends with a single vector of minimum costs from every node to the destination node. A second vector can be calculated along the way, called the *back-node vector*, containing the immediately preceding node number towards the destination; this is the most economical possible way of describing minimum paths from all nodes to a given destination. This algorithm has been improved by several authors, such as Murchland (1969) and Stoneman (1972). To minimise costly searches through the network code, Shortreed and Wilson (1968) proposed a method which used the well-known *heap-sort* algorithm for the selection of candidate nodes. This method has the rather satisfying property of processing each node only once, thus cutting computation times substantially. In de la Barra and Perez (1986) it is shown that further efficiency can be gained if the candidates are the links themselves, identified by their sequence number in the network code, rather than by their origin and destination node numbers. In this way back-nodes become back-links.

7.3.2 *Estimation of travel costs*

The above algorithm implies the evaluation of the cost of travel along each link of the network. Perceived travel costs, as mentioned, must include several elements. Firstly, there is the out-of-pocket expense or fare that operators charge to users. Then there is travel time, which must be multiplied by the value of travel time, waiting time which must be multiplied by the value of waiting time, and terminal costs, such as parking, airport taxes, storing and loading charges and so on. Since these are all user-dependent costs, the result is a variable c_{ij}^{nkpl} representing the cost of travel to user type n from i to j by mode k and link l of path p. The total accumulated cost along a path can be obtained by adding over all links that form part of it, that is:

$$c_{ij}^{nkp} = \sum_l c_{ij}^{nkpl}, \quad \forall\, l \in p \tag{7.1}$$

The value of time of each user type is a subjective concept and one of the most debatable topics in transport analysis. Essentially, its value must be obtained through calibration against real data. In the case of people, the value of time is generally assumed to be a function of income, but no such relationship can be established for other types of users, such as commodities. The fact that different users have different values of time, together with the fact that they also have access to different modes (e.g. commodities cannot travel by bus), results in the selection of different paths for each user type.

7.3.3 *Representation of public transport*

In the case of public transport, path search must include the possibility of transfers from one operator to another. In de la Barra and Perez (1986) this is solved by multiplying the dimensions of the search. In step (c) of the minimum path algorithm, candidate nodes were defined as links connected to the destination centroid; in this case a candidate is a combination of links and operators, and each combination will imply a cost. Furthermore, if along a path there is a change of operator, a transfer cost and waiting time must be added. In order to keep track of the sequence of operators that constitute a path, the dimensions of the back-link vector must be expanded also, becoming back-operator as well.

7.3.4 *Turn prohibitions*

The algorithm, so far, assumes that users or operators can move freely from one link to another if they are connected. However, it is very common, particularly in urban areas, that turn prohibitions restrict some movements. For example, in the network of figure 7.3, all left turns could be prohibited along the main avenue, a common practice in traffic management to ease the flow of vehicles in central areas (right turns in Britain). This will affect intersections 2 and 3. Turns such as 1–2–5 or 4–3–8 are no longer allowed, which means that in order to reach 5 from 1 there are only two possible paths:

1–2–3–8–7–2–5 and 1–2–7–8–3–6–5.

The minimum path algorithm explained so far cannot represent the first of these paths, because it implies the repetition of node 2. Nodes are never repeated because they are processed only once. On the other hand, if they were allowed, not only the path described above would become possible, but all sorts of absurd loops.

Therefore, the standard method of representing turn prohibitions is through the expansion of nodes. Each node representing a street intersection where turn prohibitions occur is divided into several nodes with virtual or fictitious links connecting them. Figure 7.5 shows the minimum expansion of nodes that is required to represent left-turn prohibitions at intersections 2 and 3. For example, node 2 has been divided into nodes 2a, 2b, . . ., 2h. Note that even if prohibitions only occur at intersections 2 and 3, it is necessary to expand all intersections. If, say, intersection 7 was kept as a single node, the algorithm would find that the shortest path from 1 to 5 was 1b–2b–2h–7–2g–2d–5, which would be wrong.

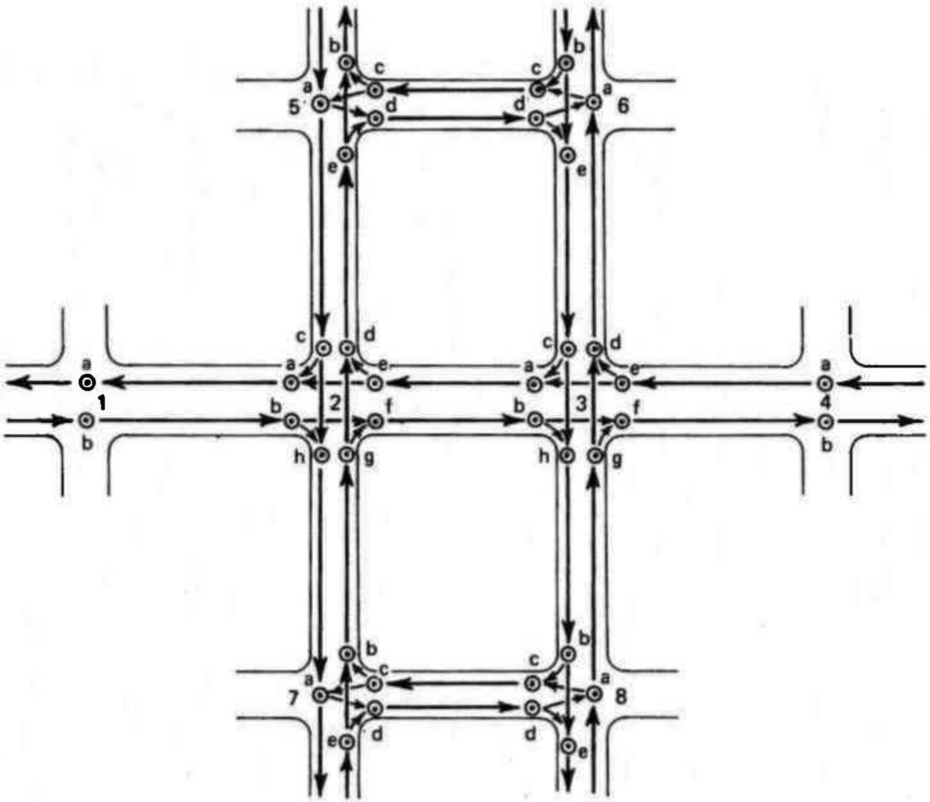
This way of representing turn prohibitions has a number of problems, which can be summarised as follows:

- The design of the expanded network is a difficult and time-consuming process, requring in some cases considerable skill on the part of the analyst.
- The addition of virtual nodes and links can increase the total dimension of the graph substantially.
- The method is error prone because the virtual nodes and links can be used by the algorithm in unexpected ways (as in the example above).
- Even if the analyst wishes to expand those intersections where turn prohibitions occur, he is forced to expand neighbouring intersections as well in a domino fashion.
- The results of the model contain a large amount of unnecessary information (virtual links).

The real problem is that the transport network is not transformed into a graph in the most convenient way. The approach adopted so far, where street segments are represented as the edges (links) of the graph and street intersections as the vertices (nodes) of the graph, can be called the *intersection-based representation*. In de la Barra and Perez (1986) a different approach is pursued, called the *reverse graph representation*.

Figure 7.6 shows the way in which the same simple network should be coded in the proposed method, including the left-turn prohibitions at intersections 2 and 3. In the code itself, each street segment is represented in the traditional way: intersections are numbered, and each street segment is described by its origin and destination intersections, except that now turn prohibitions are specified for each link. For example, street segment 1 goes from intersections 1 to 2, but cannot turn towards 5. As can be seen, the number of links and nodes in the network code remains the same as in figure 7.3 where no turn prohibitions were indicated. Prohibitions, however, are indicated in a simple way, and no virtual links or nodes are required.

Figure 7.5. The method of expanded nodes for representing turn prohibitions in a transport network

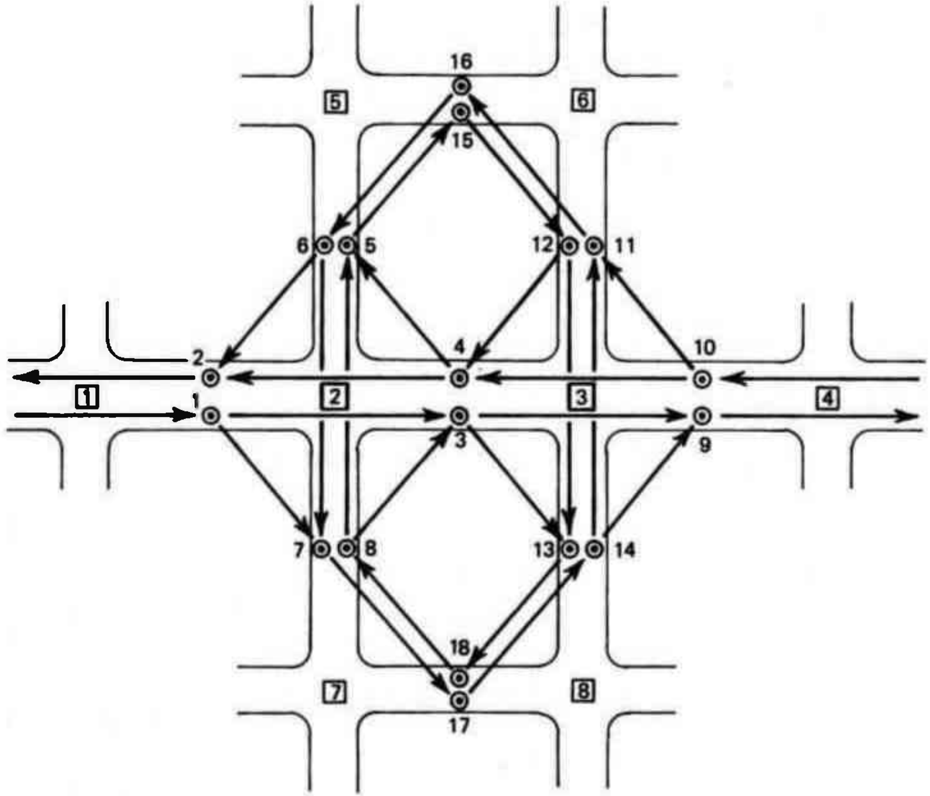


Network code

For left-turn prohibitions at intersections 2 and 3

Origin node	Destination node	Origin node	Destination node
1b	2b	5a	5d (*)
2a	1a	5c	5a (*)
2b	2f (*)	5c	5b (*)
2b	2h (*)	5d	6d
2c	2a (*)	5e	5b (*)
2c	2h (*)	5e	5d (*)
2d	5e	6a	6c (*)
2e	2a (*)	6b	6c (*)
2e	2d (*)	6b	6e (*)
2f	3b	6c	5c
2g	2d (*)	6d	6a (*)
2g	2f (*)	6d	6e (*)
2h	7a	6e	3c
3a	2e	7a	7d (*)
3b	3f (*)	7b	2g
3b	3h (*)	7c	7a (*)
3c	3a (*)	7c	7b (*)
3c	3h (*)	7d	8d
3d	6a	7e	7b (*)
3e	3a (*)	7e	7d (*)
3e	3d (*)	8a	3g
3f	4b	8a	8c (*)
3g	3d (*)	8b	8c (*)
3g	3f (*)	8b	8e (*)
3h	8b	8c	7c
4a	3e	8d	8a (*)
5a	2c	8d	8e (*)

No. of links: 54
No. of nodes: 40
(*) Virtual links



Network code (external representation)
For left-turn prohibitions at intersections 2 and 3

Origin intersection	Destination intersection	Turn prohibition to intersection:
1	2	5
2	1	–
2	3	6
3	2	7
2	5	–
5	2	3
2	7	–
7	2	1
3	4	–
4	3	8
3	6	–
6	3	4
3	8	–
8	3	2
5	6	–
6	5	–
7	8	–
8	7	–

No. of links: 18 external, 24 internal
No. of nodes: 8 external, 18 internal

□ External intersection-based representation

⊙ Internal link-based representation

Figure 7.6. Reverse graph representation of a transport network

When the network code is entered to the path-building algorithm, the reverse graph representation of the network is built, where now the street segments become the vertices of the graph, and its edges (links) represent their connectivity. The resulting reverse graph is also shown in figure 7.6.

Each edge in the internal reverse graph represents an acceptable connection to a *source* street segment. An acceptable connection is defined as another connected street segment whose destination node is not included in the list of prohibitions. The rest of the algorithm remains largely unchanged. If steps (a) to (d) are applied to the reverse graph to determine the minimum path between 1 and 5, the resulting path will be 1–3–13–18–8–5 in figure 7.6. It is evident that the algorithm will have no problems in finding such a path, because no vertices are repeated. The algorithm can easily translate its internal representation back to user-defined street intersections. To the user, in fact, the above path becomes, in terms of street intersections, 1–2–3–8–7–2–5 where node 2 is repeated. A loop, however, cannot occur, because it would imply the repetition of a vertex in the internal representation.

In the case of public transport, turn prohibitions will apply to vehicles but not necessarily to passengers, because they can transfer from one vehicle to another. The algorithm must allow for this circumstance, but add transfer costs and waiting times if the instance occurs.

7.3.5 *Multi-path search*

A path search algorithm must be viewed as an analytical way of determining the options available to users. From this point of view, it is quite clear that users will not consider a single option to travel from an origin to a destination, but rather several options. This implies that the path-building algorithm should determine the first n paths instead of just the least cost one, where n is a given parameter.

The way to modify the algorithm to include multi-path search is, again, a matter of adding another dimension to the cost vector and to the back-node and back-operator vectors. Instead of one cost label where the minimum cost of travel to a particular destination is recorded, n labels are required. Also n back-node registers are required, because the sequence of nodes in each path is different. Similarly, n back-operator labels are required. Apart from the increase in the number of dimensions, the rest of the algorithm remains the same.

7.4 Capacity restriction

The calculation of travel cost along a path was described in the previous section as an intrinsic component in the determination of which paths are considered by users as their travel options. The cost perceived by users at a path level, it was mentioned, has a number of components – among others, travel time and waiting time. In the first iteration of the

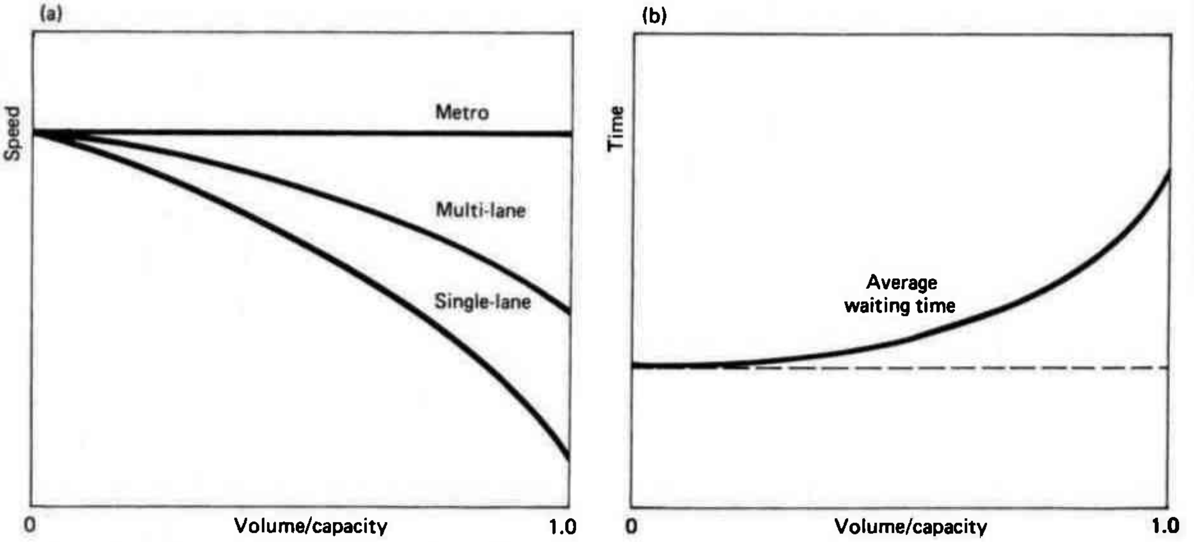


Figure 7.7. Reduction of speed in different types of links as the volume of flow increases

model, it must be assumed that the network is empty, and hence travel time is determined by free flow speeds along each path.

Once free flow perceived costs have been estimated, the transport model will calculate demand, separate it by modes and finally assign it to paths. At this point demand must be confronted with transport supply available in each link along paths, increasing travel time accordingly. The way in which travel time increases depends on a number of factors, mainly the characteristics of the link itself and of the operator.

Figure 7.7(a) shows the effect of different road types on the reduction of speed as the volume of traffic nears capacity. Speed along a multi-lane motorway will remain almost constant as traffic increases, reducing speeds only near the $V/C = 1.0$ limit. A single-lane road, because of higher vehicle interference, will reduce speed more quickly. Vehicles along a metro line will not reduce speed, because the number of vehicles is fixed.

In the case of public transport, whether of people or commodities, two kinds of reductions must be taken into consideration: vehicle reduction and user reduction. Vehicle speed reduction follows the same rules explained above, but passengers of public transport systems or commodities on cargo systems have waiting times that can increase as the demand for such systems nears capacity. This is shown in figure 7.7(b).

7.5 Trip generation

In section 7.2, when describing the integrated land use and transport model, the trip generation model was one of the two fundamental interfaces between the activity system and the transport system, together with generalised composite cost. The ideal trip generation model, it was argued, is a function that transforms the functional flows from the activity model into actual trips, taking into consideration the generalised cost of travel. In other words, it must be treated as an elastic demand function, such as the one shown in figure 7.8.

As can be seen, the demand function for transport is similar to the demand function for floorspace in figure 6.3, so that there is no reason to adopt a form different to equation (6.2) which in this case becomes:

$$T_{ij}^{n} = X_{ij}^{n}[a^{n} + b^{n}\exp(-\beta^{n}c_{ij}^{n})] \qquad (7.2)$$

where T_{ij}^{n} is the total number of trips generated from i to j by socio-economic category n, and X_{ij}^{n} are the functional flows produced by the activity model. Parameter a^{n} is the minimum number of trips activity n must perform, and $a^{n} + b^{n}$ is the maximum. The number of trips decays from $a^{n} + b^{n}$ to a^{n} exponentially, with a slope regulated by β^{n}, as the generalised composite cost of travel, c_{ij}^{n}, increases.

Elasticity in travel demand prediction is of paramount importance. Any new transport facility introduced to the system, with few exceptions, will *induce* new trips. For example, traffic counts carried out before and after the introduction of a thirteen kilometre metro line in Caracas showed an increase of 160% during peak hour in a particular corridor. The opposite case is also frequent: many times planners observe the growth of traffic on a particular road, project it and conclude that in, say, twenty years' time the capacity of that road must increase three times to cope with demand; such a catastrophe might never take place, because the small capacity of the road inhibits trip generation. Another good example was the completion of the M25 motorway in London; planners had argued that this road would divert traffic from the central areas and that it would take many years for demand to reach capacity; surprisingly, it only took fifteen days, and the central area has shown no signs of improvement. Planners' misjudgement became a popular issue in the local press. If only they had had an elastic trip generation model such as equation (7.2)!

But in spite of the important implications of elasticity in trip generation, it is surprising that very few transport models are able to estimate it. To the author's knowledge, the only exception is Echenique, Geraldes and Williams (1977), where elastic trip generation rates are solved in

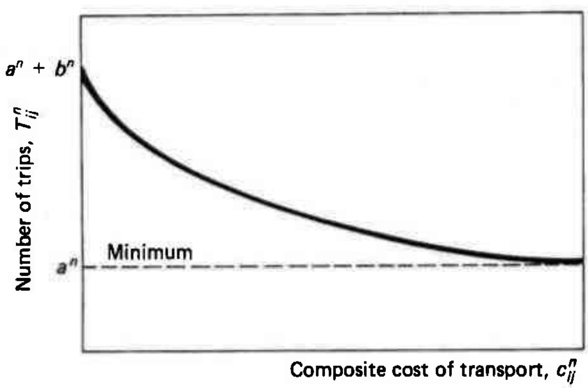


Figure 7.8. Elasticity in the generation of trips

the activity model: households equilibrate an Alonso type budget equation with the classic three elements: land, transport, and all other goods and services. From here, the amount of money a household is willing to spend on transport is calculated, and the information is transferred to the transport model, where money is converted into trips. This solution is indeed an elegant one, but difficult to calibrate in practice with available data. The other problem is that trip generation varies not only with income and location, but also depends on the hours of the day. For instance, for 24 hour trips, revealed elasticities are generally lower than for peak hour trips. In the example of Caracas, many of the new trips were performed at off-peak hours before the introduction of the metro. Also, in the case of commodities, there is no budget constraint equation.

7.6 Modal split

In transport modelling in general, a wide variety of theories, algorithms and analytical tools are used to solve otherwise similar problems. Modal split is, however, an area where most of the research coincides in that the multinomial logit model is the most appropriate. In this and previous chapters it has been argued that it is not only appropriate for the representation of the modal split phenomenon, but for the representation of all choice situations, from the location of activities to route assignment. In fact, the modal split case has been used in previous chapters as an example of a typical choice situation.

The general random utility based model of equation (4.10) is directly applicable to the modal split problem, which in this case adopts the form:

$$T_{ij}^{nk} = T_{ij}^{n} \frac{\exp(\beta^n c_{ij}^{nk})}{\sum_k exp(\beta^n c_{ij}^{nk})}, \quad \forall k \in K^n \qquad (7.3)$$

The expression $\forall k \in K^n$ denotes summation over all modes k belonging to the set K^n of modes available to category n. In the case of commodities, K^n is restricted to cargo modes, such as lorries or freight trains. In the case of passengers, however, it raises the problem of car availability, a subject well covered in Supernak (1983). Each passenger category will have a particular car availability rate; those with no car available will be assigned directly to public transport modes, while those with a car available will be subject to modal split.

Car availability, as argued by Supernak (1983), is a better concept than the traditional car-ownership. It is very common that people do not own a car, but still have one available, as in the case of company cars. It is also often the case that there is one car in the family, which is used by one of the members, leaving the rest of the members without a car available, and there are many cases where there is more than one car in the family.

Many authors point out that the modal split model of equation (7.3) can lead to errors because of attribute correlation among options. The classical example is the well-known 'red-bus–blue-bus' conundrum (Mayberry, 1970), which can be explained as follows. Consider a simple case where there are two modes available to travel between a particular origin and destination: $k = 1$, car; $k = 2$, bus. For simplicity, assume that both modes have equal costs and times, so that they are indifferent to users. In this case, equation (7.3) will result in an equal number of trips in each mode, which in terms of probabilities is $P^{car} = P^{bus} = \frac{1}{2}$. If now the bus mode is divided into two sub-modes by merely painting them red or blue, modal split will now take place between three modes: $k = 1$, car; $k = 2$, blue bus; $k = 3$, red bus. Since users still perceive the same cost from the three alternatives, probabilities will now become $P^{car} = P^{blue} = P^{red} = \frac{1}{3}$.

This is obviously a fallacy, because colour should produce no effect on the results. The solution proposed by Williams (1977a, b) is to arrange the modal split equation in a hierarchical fashion, because users will consider both types of buses as very close options. In this case, the probability of choosing a blue bus will be equal to the probability of choosing buses of any colour over cars, multiplied by the probability of choosing blue buses over any colour, that is: $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$, which is the correct answer.

The solution given in section 7.3 is equivalent to Williams' because different types of buses were considered as public transport operators within a single public transport mode. In this case, the second hierarchical level has been built into the path search algorithm.

7.7 Route assignment

Trip assignment is the process by which trip matrices by mode and user type, T_{ij}^{nk}, are transformed into the number of trips that use each link of the transportation network, T^{nl}. To perform this task, there is a wide variety of methods. Wardrop (1952) identifies two criteria on which the various assignment methods can be based:

(a) The first criterion uses the concept of *average travel time*, and assumes that users act selfishly, considering only their individual travel times in making the route choice decision.
(b) The second criterion is based on the concept of *marginal travel time*, and assumes that users are aware of the way their individual route choice influences the overall travel time of the system.

Average and marginal times are represented in figure 7.9. Average travel time represents the time actually experienced by individual users, and will tend to increase as the volume of traffic along a link increases. Marginal travel time represents the increase in the aggregate travel time when one new user enters the traffic stream.

If the first principle is assumed to prevail, then the method can be called a *user-optimised* assignment. If, on the other hand, Wardrop's second criterion prevails, the method can be called a *system-optimised* assignment. This second type of assignment will lead to a lower total cost, but it will be less realistic. Reality probably lies somewhere between

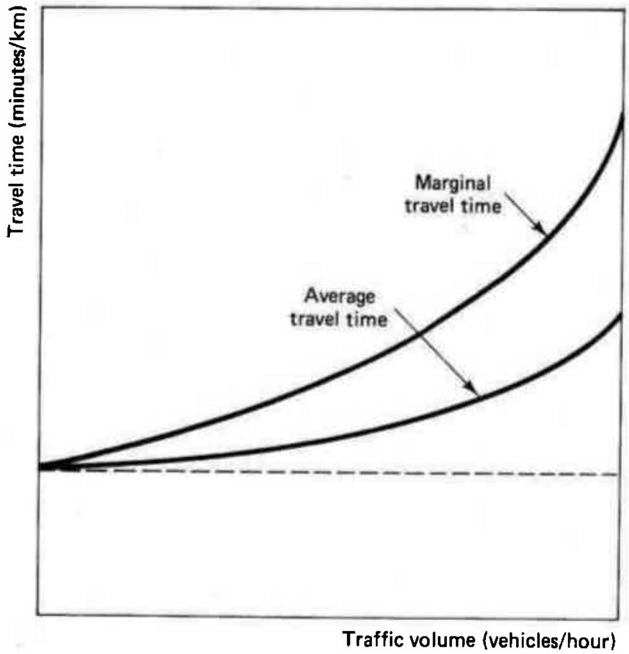


Figure 7.9. Average and marginal travel time

the two criteria; it can be argued that a random utility based assignment model produces results which are also halfway between user and system optimisation. It is sometimes proposed that both methods should be used: first use the system-optimised method, and then devise policies to influence user-optimised behaviour to bring it close to the system-optimised pattern.

The different asignment techniques that are available in the literature have evolved from the most simple to more elaborate techniques, as both computer hardware and algorithms have been developed. In the following sections, the main methods are briefly reviewed.

7.7.1 Single-path assignment

This is the most simple and cost-effective solution. The method proceeds by identifying the least cost route from given origin and destination pairs, and assigns the total number of trips to that route, on an all-or-nothing basis. The method clearly falls into Wardrop's first criteria. It is easy to imagine that if this method is applied to a dense network, the results will be unreal, because traffic will be assigned only to those links that lie within at least one minimum path; it will also overload links that happen to lie within many minimum paths. For these reasons, this method is hardly ever used today.

7.7.2 Incremental assignment

Incremental assigment represents an attempt to solve some of the deficiencies of the single-path assignment method, by introducing a capacity restriction procedure. It also falls within Wardrop's first criterion. The method calculates assignment in an iterative way. In the first iteration, free flow speed times are calculated and used as a basis for minimum path search. A small proportion, say 30%, of the trip matrices are then assigned to the minimum path on an all-or-nothing basis. Travel times are then re-estimated with a capacity restriction function, and a new set of minimum paths is searched for. In a second iteration, another, even smaller, proportion of the trips are assigned to the new paths, say 20%. Subsequent iterations are performed, until 100% of the trips have been assigned.

This method, included in many standard transportation computer packages, such as the UT PS package (UMTA, 1977), has the advantage of combining capacity restriction with the operational efficiency of single-path assignment. It retains, however, some of the problems of its predecessor and adds some of its own. The first problem is that the method has no heuristical basis, because in reality transport users do not behave in an incremental way. The increments are mere pre-defined guesses with no theoretical basis. The second problem is that, if the network is not highly congested, such as a medium-size town or a regional system, it may well be that assignments with the incremental algorithm are indistinguishable from those obtained with the single-path assignment, because the set of minimum paths could be the same in all iterations. From a theoretical point of view, the method assumes that users will always want to travel along the minimum path, and only divert from it if forced by congestion. In chapter 4, when discussing random utility theory, it was shown that users disperse around the minimum path, mainly because of the aggregation process involved. In fact, incremental assignment tends to produce unreal results, with many links unloaded because they never managed to be included in any particular minimum path.

7.7.3 *Equilibrium assignment*

This method is based on the principle that users choose between alternative routes in such a way that, in equilibrium, no individual user will be able to improve its utility by switching to another route. Hence, Wardrop's first principle is at the very heart of the method. Equilibrium is achieved when all users are in their individual minimum cost paths, or, in other words, when travel time is equal in all paths connecting an origin–destination pair. The method is well described in Kanafani (1983).

Assume, for simplicity, that there are two possible paths connecting an origin i to a destination j. For each one of these routes, capacity restriction imposes an increasing cost or time as a larger number of users decide to join that path. Figure 7.10 shows the way in which the two cost functions grow as the number of users increase; the two cost functions have been arranged so that one is plotted from left to right and the other from right to left, and the horizontal axis has been arranged such that it matches the total number of trips to be assigned, T_{ij}^{n}. These

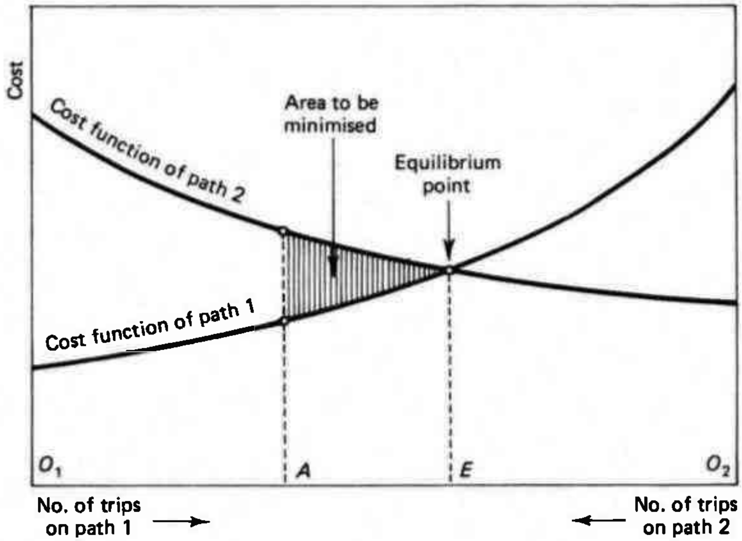


Figure 7.10. Cost functions in the equilibrium assignment algorithm

two functions will cross at point E where the perceived cost of the two paths is equal. At any other point different to E, a certain number of users would benefit by switching from one road to the other. The problem can be formulated in mathematical programming terms by establishing as an objective function the minimisation of the area under the cost curves (shaded area in the diagram).

Le Blanc, Morlok and Pierskalla (1975) developed an algorithm that can solve the equilibrium assignment problem for large networks efficiently. The method proceeds iteratively in the following steps:

(a) Assume an initial distribution of trips to paths on each link of the network and compute the corresponding costs.
(b) With these costs, find the minimum path.
(c) Assign all trips to the set of minimum paths.
(d) Calculate a weighted average between the initial assignment in (a) and the one obtained in (c); weights are obtained heuristically so as to minimise the area under the objective function.

Calculation goes back to step (a) where the initial estimate is replaced by the resulting assignments of step (c). The algorithm has been shown to converge efficiently, as long as the cost functions are convex.

Equilibrium assignment is available in many computer packages, such as the UT PS (UMTA, 1977). This algorithm can be criticised from two main points of view. From a theoretical point of view, it assumes user average cost optimisation, which has been shown to be unreal. As a result, if the network is not heavily loaded, the algorithm will tend to overestimate flow along main roads and underestimate minor ones, a behaviour similar to the incremental assignment algorithm. From a practical point of view, as discussed in section 7.3, cost functions are not always convex.

7.7.4 *Probabilistic route choice*

The main assumption of this method is that traffic will not always flow along the minimum cost route, but will spread across alternative routes instead. One of the first pioneering efforts in this direction is due to McLaughlin (1966), who proposed a multi-path assignment algorithm with capacity restriction based on *path enumeration*. In general terms, it does not load traffic to the minimum path only, but to the N-shortest paths instead. Calculations proceed in the following steps:

(a) Compute all paths between each origin–destination pair that have costs less than a specified multiple, P, of the least cost path ($P > 1$).
(b) For each origin–destination pair, construct a sub-graph in which each link represents one of the candidate paths.
(c) Use linear graph theory to determine the flows on each link of the sub-graph. Each alternative path will receive flow inversely proportional to its impedance.
(d) Apply capacity restriction to each link to determine restricted flow speeds and go back to step (a) unless convergency has been achieved.

As can be seen, this early attempt of developing a multi-path probabilistic assignment method is very much along the lines proposed in section 7.3. The algorithm, however, found many operative problems, which were probably due to the fact that at the time of its development technology could not cope well with its computing and storage requirements, and that efficient multi-path search algorithms were not yet available.

Dial (1971) came to the rescue with a paper suggestively titled 'A probabilistic multi-path traffic assignment model that obviates path enumeration'. The method was proposed as a way of dealing with large networks economically, and is currently available in standard packages, such as UT PS (UMTA, 1977).

The method proceeds as follows. Consider the example of figure 7.11(a) where, say, 100 trip-makers intend to travel from centroid 1 to centroid 2. In the network there are also nodes 3 to 9. Assume that a minimum path algorithm has been used to calculate minimum travel costs from every node to destination node 2. Contrary to the minimum path algorithm explained in section 7.3 above, the assignment algorithm will proceed from the origin towards the destination.

Let $c(n,2)$ be the minimum cost of travel from a node n to the destination node 2. The algorithm begins by selecting possible candidate nodes from the origin. All nodes connected to origin 1 are potential candidates, but the following condition is imposed for a candidate to be accepted:

$$c(n,2) \leqslant c(i,2), \tag{7.4}$$

where n is the candidate and i the origin. Condition (7.4) can be called

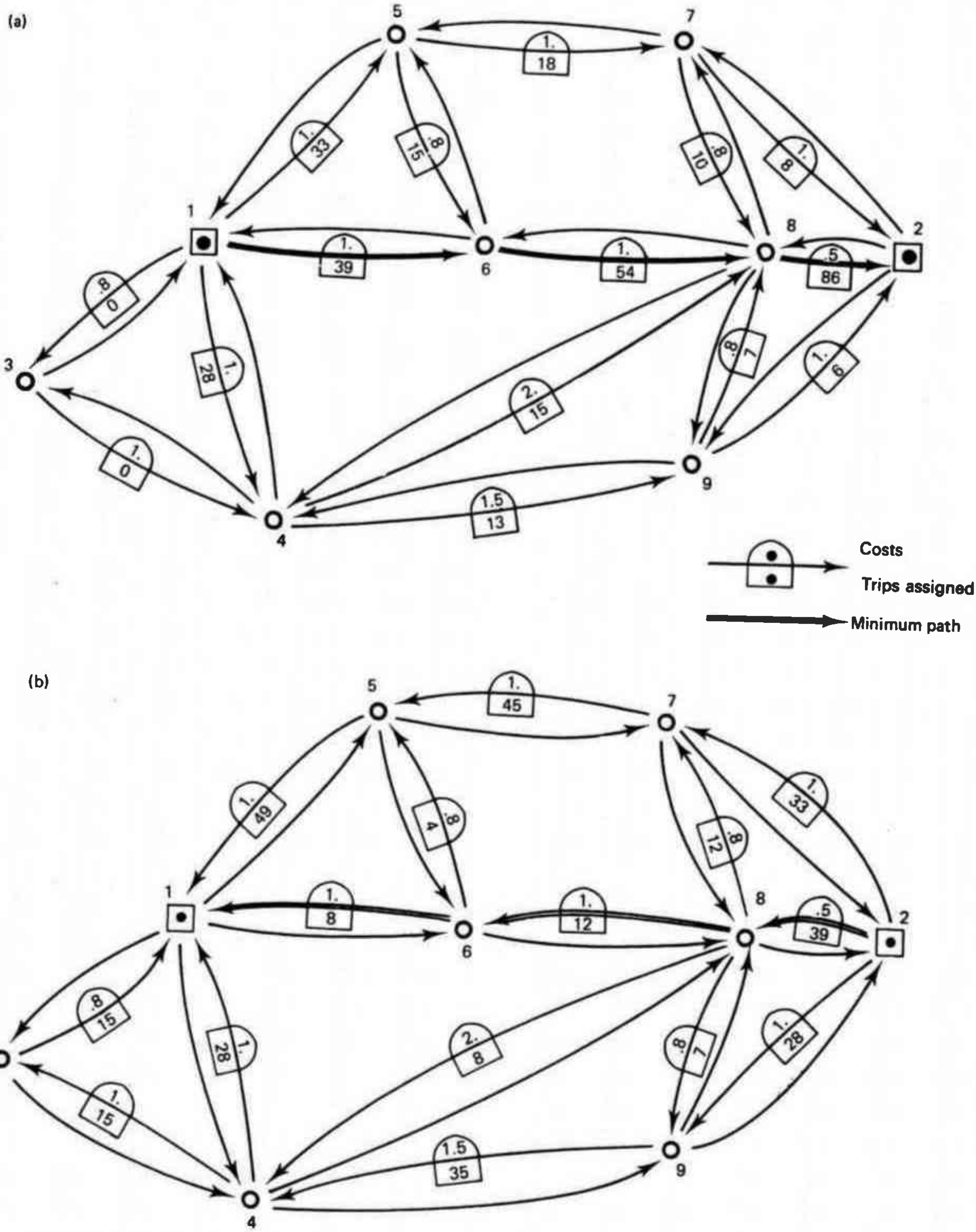


Figure 7.11. An example of Dial's trip assignment algorithm: one hundred trips (a) from 1 to 2, and (b) from 2 to 1

the *approaching condition*, because it states that node *n* is an acceptable candidate if it brings the trip-maker closer to its destination. In the example of figure 7.11(a), there are three candidates from origin 1: 4, 5 and 6. Node 3 is not acceptable because the cost of travelling from 3 to 2 is 3.5, whereas the cost of travelling from 1 to 2 is 2.5.

Once the candidates have been determined, traffic is assigned probabilistically to them, labelling each node with the resulting trips. For the sake of the argument, the exact form of the probability function is irrelevant, the obvious choice being a multinomial logit model, but for simplicity and to allow the reader to check the results, the simplest possible form has been adopted. Table 7.1 shows the resulting costs, probabilities and assignments.

Each of the candidates is then taken as the origin. For example, node 5 with a label of 33 is now the origin. From node 5, two candidates can be accepted: 6 and 7. Similar calculations are performed until all trips reach destination node 2. Program DIAL can be used to experiment with the algorithm.

The main attraction of the method is its computational efficiency, in spite of the fact that it performs multi-path probabilistic assignment; hence its popularity. It has, however, some important deficiencies, the main one being the symmetry of the resulting assignments. To understand this problem, consider the opposite case to the one described above, that is, 100 trips travelling from node 2 to node 1, assuming that all links have exactly the same characteristics in both directions. The procedure described above is used to estimate the assignments; the results are shown in figure 7.11(b). When figures 7.11(a) and 7.11(b) are compared, it is immediately apparent that the resulting assignments are dramatically different; only two links show the same results (1–4 and 9–8), two links show zero flow in case (a) (1–3 and 3–4), and the rest of the links show asymmetries from -77.8% to 366.7%.

From the above results, it is clear that there is a fundamental limitation to the algorithm. There is no reason to assume that trip-makers will choose different paths when travelling in opposite directions, given a perfectly symmetrical network. The only acceptable sources of asymmetry are asymmetric trip matrices to be assigned and/or asymmetric links in the network. It is easy to demonstrate that such asymmetries do not exist in reality. The conclusion is that the strong asymmetries

Table 7.1.

Node	Costs	Probabilities	Assignment
4	$c(4,2) + c(1,4) = 3.5$	$3.5^{-1}/1.019 = 0.28$	$\times 100 = 28$
5	$c(5,2) + c(1,5) = 3.0$	$3.0^{-1}/1.019 = 0.33$	$\times 100 = 33$
6	$c(6,2) + c(1,6) = 2.5$	$2.5^{-1}/1.019 = 0.39$	$\times 100 = 39$
$3.5^{-1} + 3.0^{-1} + 2.5^{-1} = 1.019$			Total = 100

shown in figure 7.11 result from the logic of the algorithm, not from any behavioural considerations.

In de la Barra and Perez (1986), it is shown that the main source of asymmetry in Dial's algorithm is the approaching condition (7.4), because it is not independent of the topology of the network, and that unfortunately the condition cannot be removed. It is also shown that an algorithm based on path enumeration similar to McLaughlin's (1966) with the improved path search methodology described in section 7.3 above can be both efficient and have perfectly symmetrical properties. The final form of the proposed algorithm is:

(a) Compute the first *n*-paths connecting each origin–destination pair, mode and user type, storing back-nodes and back-operators.
(b) Assign trips to paths with a multinomial logit model of the form:

$$T_{ij}^{nkp} = T_{ij}^{nk} \frac{\exp(-\beta^n c_{ij}^{nkp})}{\sum_p \exp(-\beta^n c_{ij}^{nkp})} \tag{7.5}$$

(c) Once all link loads have been calculated, perform capacity restriction to calculate the resulting restricted flow speeds.
(d) Finish if convergence has been achieved; otherwise go back to step (a).

The proposed algorithm meets some important requirements. Apart from its symmetrical properties, it can be applied to commodities or personal trips, and in the latter case, it can deal with automobile and public transport alike, adequately representing transfers between operators, waiting times, prohibited turns and other factors. It is, however, subject to the general criticism against multinomial logit based models related to attribute correlation.

Attribute correlation in assignment models can be explained by yet another example, similar to the 'red bus–blue bus' conundrum, this time called the 'hole on the road' problem, described in figure 7.12. In figure 7.12(a), node *i* is connected to node *j* by two alternative paths ($p = 1,2$) which, for simplicity, will be assumed to have equal costs. In this case, the assignment model of equation (7.5) will determine that trip-makers will be indifferent as to which of the two they should

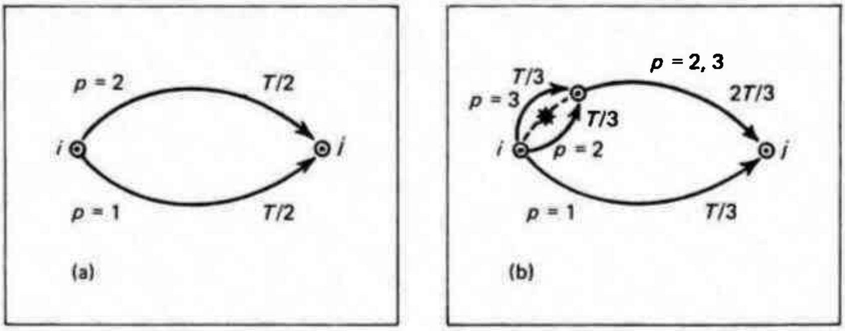


Figure 7.12. Example of the 'hole on the road' to explain attribute correlation problems in the assignment model

choose. Probabilities in each path will be equal, so that if T is the total volume of traffic from i to j, then each path will carry $T/2$. One day, a big hole opens in one of the paths, and the King orders the construction of two diversions to avoid it, creating a new situation described in figure 7.12b, and asks his transportation planner to predict traffic distribution. The planner correctly assumed that the diversions would add a negligible cost, and his path search algorithm found that there would be three paths ($p = 1,2,3$) of equal costs connecting i to j. His head was cut off when he applied equation (7.5) to estimate future traffic and found that each path would carry $T/3$, and that the problematic road would carry more traffic than before ($2T/3$).

The problem is that paths 2 and 3 are highly correlated, and hence they are considered as a single option by users. To be consistent with discrete choice theory, the multinomial logit based assignment model should be arranged in a hierarchical fashion. The network would have to be coded explicitly recognising an aggregate level and a disaggregate level. In the first hierarchical level, there would be only two paths, whether or not the diversion had been built. In both cases, probabilities in each path would be $T/2$. At a second hierarchical level, path 2 would have two sub-paths, and each sub-path would have a probability of $T/2 \times 1/2 = 1/4$. A hierarchical assignment model would have saved the planner's life.

This is another example showing that the problem of attribute correlation can be solved through the hierarchical structuring of models. In the case of traffic assignment, though, this solution is still difficult to achieve in practice, since efficient algorithms are not yet available, and remains as an area of research and development of great potential. For the time being, the analyst must be careful when coding the network to avoid problems of this kind.

7.8 Composite costs and consumers' surplus

The last point that remains to be seen is the exact form in which composite costs of travel must be aggregated. It is mainly a matter of applying equation (4.11) to each link in the decision chain, as explained in section 7.2 when describing the general structure of the integrated land use and transport model.

The rule is that composite costs must be computed following the decision chain backwards. In this case, calculation must begin with travel cost at path level, then at mode level, and finally the composite cost of travel for each origin–destination pair and user type. The calculation of costs at path level has already been explained in section 7.3, generating a set of c_{ij}^{nkp} matrices. These costs must now be aggregated with respect to paths, to obtain the composite cost of travel from an origin i to a destination j by mode k, over all possible paths, c_{ij}^{nk}:

$$c_{ij}^{nk} = \frac{1}{\beta^n} \ln \left[\sum_p \exp(-\beta^n c_{ij}^{nkp}) \right] \tag{7.6}$$

where β^n represents the assignment parameter in equation (7.5). The resulting c_{ij}^{nk} values are then used by the modal split model of equation (7.3). The next step is to aggregate these costs over all modes, to obtain the generalised composite cost of travel between zones by user type, c_{ij}^n, based on the results of (7.6):

$$c_{ij}^n = \frac{1}{\beta^n} \ln \left[\sum_k \exp(-\beta^n c_{ij}^{nk}) \right] \tag{7.7}$$

where now β^n is the modal split parameter of equation (7.3). The resulting c_{ij}^n values are then used by the trip generation model of equation (7.2), and by the activity allocation model of chapter 6. They will also serve as a general accessibility indicator for each user type.

At this point the user benefit or consumer's surplus indicator of equation (4.12) can be introduced. There could be one indicator per choice, i.e. route choice and mode choice. However, since c_{ij}^{nk} is transferred from equation (7.6) to equation (7.7), only consumers' surplus at the mode choice level must be evaluated to avoid double counting, and in this case it would be:

$$\Delta S_{ij}^n = -\frac{1}{\beta^n} \ln \left[\frac{\sum_k T_{ij}^{nk} \exp(\beta^n c_{ij}^{nk}) ②}{\sum_k T_{ij}^{nk} \exp(\beta^n c_{ij}^{nk}) ①} \right] \tag{7.8}$$

where 2 denotes the scenario being evaluated and 1 the base case scenario against which 2 is being compared. Note that T_{ij}^{nk} has been included in the formulation to account for the difference in the number of trips in each scenario due to elastic trip generation. The set of indicators (7.8), which represent the difference in utility per trip made, must be aggregated to obtain a system-wide single value.

7.9 Conclusions

In this chapter, the main elements of the transport system and its integrated modelling to the activity system have been described. The subject is very extensive and the related literature is abundant, so that the contents of this chapter have been limited mostly to the description of the proposed models. The first two sections were devoted to the relationship between the activity and the transport systems, explaining how a linked structure can be transformed into a proper integrated structure, where activity and transport variables depend on each other. On the one hand, it has been argued that the transport system cannot be represented properly without the integration to the activity system, particularly if long-term predictions are pursued. On the other hand, if the detailed representation of the regional economy of a country or the location of activities in a city is to be useful, then it is necessary to include an equally detailed representation of the transport system.

The third section was devoted to network analysis, and to the problem of path search in particular. Transport networks are represented as directed graphs in the majority of transport models, and in this section it was demonstrated that the intersection based representation is not the most appropriate. An alternative link based representation was proposed that can solve many problems, particularly turn prohibitions, and be more efficient at the same time. It was also emphasised that the path search problem must not be kept separate from the trip assignment algorithm.

The remaining sections were devoted to other key elements in the representation of transport, such as capacity restriction, trip generation, modal split, route assignment and the aggregation of composite costs. Two main conclusions can be made. The first one relates to the theoretical consistency with which all choice elements must be treated; for this reason modal split and route assignment were treated as choice situations and the corresponding multinomial logit models were derived. If such consistency is not pursued, it is impossible to aggregate composite costs as suggested in chapter 4. The second aspect relates to elasticity in trip generation, and it was argued that the number of trips must be a function of costs. This important issue, in spite of the fact that it is well recognised in the literature, escapes most existing transport models, and makes the results of many travel demand studies unrealistic. Particular examples were given to support this view, and in the next and final chapter, further examples will be described.

8

Applications of TRANUS, an integrated land use and transport model

In this chapter, a number of case studies are presented where a specific modelling package, TRANUS, has been applied. The package is based on much of the theoretical framework described in the previous chapters. The purpose of presenting these applications is to illustrate particular instances in which the proposed theories and methods can be helpful in the solution of many real planning problems.

The chapter begins with a brief description of the modelling package, followed by a case-by-case explanation of each application. The case studies described are not the only ones where the package has been applied; they have been selected as the most representative cases in each of the following areas: urban land use planning, urban transport planning and regional transport planning. A final example related to energy evaluation is also included. The author has been involved in all case studies described here together with colleagues, particularly with B. Perez, A. Morais, M. E. Botero and P. A. Rickaby.

8.1 Brief description of the TRANUS system

TRANUS is an integrated land use and transport model that can be applied at an urban or a regional scale. The program suite has a double purpose: firstly, simulation of the probable effects of applying or implementing particular land use and transport policies and projects, and secondly, the evaluation of these effects from social, economic and financial points of view.

The TRANUS system has been developed by de la Barra and B. Perez since 1982. It is based on many of the theoretical proposals that have been presented and discussed in the previous chapters. From its beginnings, a number of design requirements were adopted:

- The system had to be simple and efficient in operation, permitting implementation on small microcomputers, but without sacrificing detail, particularly with respect to the transport system.
- In order to achieve theoretical consistency, the system had to adopt a single theoretical framework for the representation of all land use and transport phenomena, based on nested multinomial logit models.
- The system had to include at least two hierarchical levels of spatial

disaggregation, not only as a means to save computer and data resources and to allow the analyst to concentrate on particular areas of the system, but also to improve the quality of the results.

- The system had to provide the possibility of application at any level of disaggregation, from a detailed urban scale to a large regional scale; if combined with the use of hierarchies, it should be possible to apply the system to a combination of urban and regional problems.
- The land use model should be able to distinguish between a variable number of employment, population and land use types, should allow for the generation of flows between all activities, and should include an explicit representation of the property market.
- The transport model should include a number of features, such as elastic trip generation, repressed demand analysis, multi-path search and assignment, the ability to combine commodities and passengers, prohibited turns, transfers between public transport operators, integrated fares, waiting time calculations, and extensive evaluation indicators.
- The design of the programs should be flexible to ease applications to many different contexts, particularly third world situations.
- The system has to be provided with a number of complementary facilities, such as interactive data entry, calibration aids, data processing facilities, extensive evaluation programs, and facilities to report the results of the analysis.

8.1.1 General structure of the model

An explicit dynamic structure relates the two main components of TRANUS: land use and transport. The way in which land use relates to transport through time is shown in figure 8.1, where discrete time intervals are represented as t_1, t_2, t_3, and so on. The land use model estimates the location of activities in the different zones in which the study area has been divided, and equilibrates a property market. The activity allocation process results in a set of X_{ij}^n matrices representing functional flows from i to j by socio-economic sectors n. In the transport model, these matrices of flows are transformed into trip matrices T_{ij}^n by the trip generation model. In other words, functional flows determine travel demand for each origin–destination pair, and it is assumed that

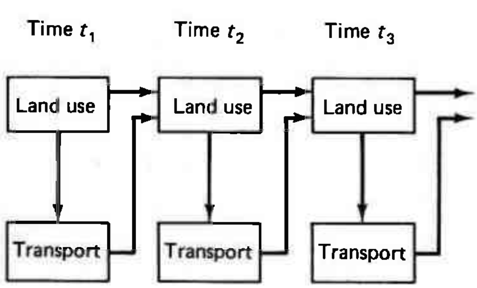


Figure 8.1. Dynamic relations in the land use and transport system

this occurs instantly in a single period of time. Thus, activities in period t_1 determine transport demand in period t_1, and so on.

The transport model, in turn, calculates the generalised composite cost of transport, which is one of the components of the utility function in the activity allocation model. Transport costs are then fed back to the land use model, and it is assumed that this does not occur instantly, but only after a time period has elapsed. When a new transport facility has been introduced, like a new motorway or a metro system that improves accessibility in particular areas, activities will take some time to react and will change their locations gradually. Hence, transport costs calculated for time t_1 influence the location of activities in time t_2, and so on. Since there are also elements of inertia in land use from one period to the next, the effects of changes in transport could well take several periods to consolidate.

8.1.2 Land use calculations

The general structure of the model is shown in figure 8.2. The starting

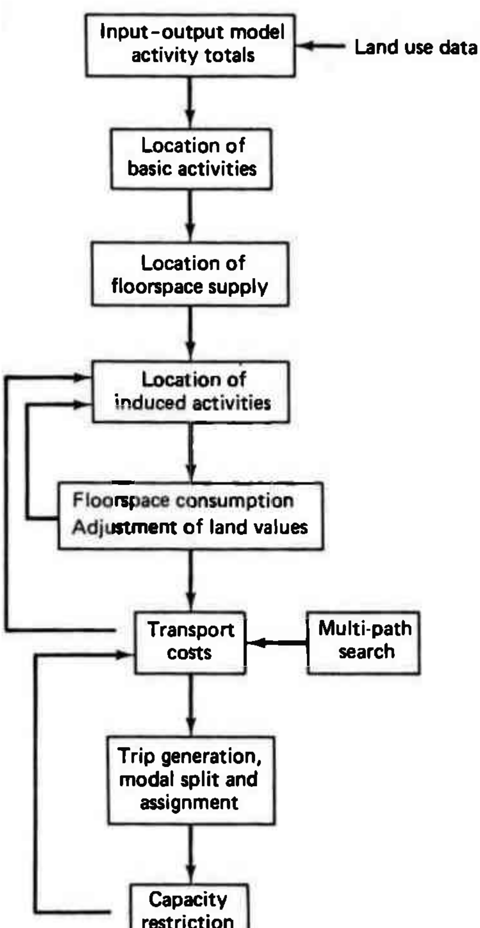


Figure 8.2. Calculation sequence used by TRANUS

point for activity location is an aspatial input–output model which distinguishes between basic final demand sectors and induced sectors. The number of sectors within each group can be freely determined by the analyst. Given exogenously defined total basic employment and the set of technical coefficients a'''', the model determines totals of induced employment and population.

The location of the increments (positive or negative) of basic employment is estimated, taking into account spatial hierarchies if they have been defined. The analyst can exogenously assign parts of the increment to particular zones; such would be the case if the location of a particular industry was part of the set of policies.

This is followed by the location of the increments of floorspace. The model assumes that any growth in activities is accompanied by a proportional growth in the supply of floorspace. It will first determine the magnitude of this growth and then will estimate its probable location hierarchically. If, however, growth in activities is negative, the model will assume no growth in floorspace. The utility functions for the location of increments in floorspace include the following elements: floorspace in the previous time period, equivalent land values, and potential floorspace. Equivalent land values differ from ordinary market values in that the former considers the intensity of land use, that is, the area of floorspace that can be built per unit area of land, which is, in turn, determined by planning regulations. Whenever land rent is taken into consideration, the model uses equivalent land values, but once the land market has been brought to equilibrium the model translates these values back to market values. Potential floorspace, on the other hand, is the area of floorspace that can be added to a zone, calculated as the difference between the maximum floor area allowed by regulations, and the floor area that existed in the previous time periood. As in the case of basic employment sectors, the user can predetermine the location of parts of the increment, to represent public housing projects (positive) or demolitions (negative).

The next stage in the sequence of calculations is the hierarchical location of induced activities, and estimation of the resultant floorspace consumption in each zone as a demand function of land values. Consumption is compared to supply in each zone, and land values are adjusted accordingly. This starts an iterative process which ends when supply and demand are approximated in every zone. Induced activities generate flows, and these are calculated together with the spatial allocation process. The analyst may specify which activities participate in the property market; for instance, if educational employment has been defined as one of the induced socio-economic sectors, the analyst may decide that such activities occupy their own premises, and hence do not participate in the market; similarly, marginal or squatter population may be left out.

8.1.3 Transport calculations

In the simulation of the transport system, the first task of the model is to search for the first n-paths connecting origin–destination pairs by mode. Each mode, in turn, can be divided into several transport operators, assuming that a trip-maker can freely transfer from one operator to another within a mode. For example, three modes can be defined: cargo, public transport and private transport; in turn, cargo can be divided into light and heavy trucks and freight trains, and public transport can be divided into buses, metro and passenger trains. In this example, commodities may transfer from light trucks to trains and back to heavy trucks, or any other combination. Public transport passengers may take a bus to a metro station and then take the metro to a passenger train.

The transport network is defined as a set of interconnected links. Each link is described in terms of its main characteristics, such as link type, link length, physical capacity, capacity of each public transport operator and prohibited turns. Link type, in turn, defines a number of characteristics of each link, such as speed, circulation charges (tolls), maintenance costs and the administrator in charge of the link. At the end of the process, each path is represented as a sequence of operator and link pairs that connect an origin centroid to a destination one. When along a path there is a change of operator, a transfer is assumed, adding a transfer cost and a waiting time to the cost of the path.

Once all paths have been found, the first step of the iterative transport calculations is to estimate the cost of travel along each path, the composite cost by mode and composite generalised transport cost. The cost of travel along each path is recalculated in every iteration to account for changes in travel and waiting times due to congestion.

The activity model produces a set of matrices called functional flows, which the transport model must turn into actual trips at particular hours of the day (peak hour, 24 hours, etc.). The analyst can mix functional flows in different proportions, resulting in transport categories. For instance, socio-economic categories such as agriculture, mining and industry can be transformed into transport categories such as light cargo, heavy bulk, liquid bulk, and so on. Trip generation determines the number of trips that will be made from an origin to a destination by a particular transport category, as an elastic function of the corresponding generalised composite costs. Elasticity in travel demand means that for a given functional relationship, more trips will be made if there is a reduction in the cost of travel. As a result, in each iteration, the number of trips will be reduced as congestion builds up. The difference between the number of trips in the last (equilibrium) iteration and the first iteration will represent the number of trips that are not made because of congestion, and this is called *repressed demand.*

Trips for each category are then distributed to modes with a multinomial logit model where the utility function is determined by the composite cost of travel by mode. Distribution is made over all modes available to each category. For instance, a trip category representing commodities will only choose from cargo modes, such as trucks, ships, and so on. Modal choice for passenger categories is restricted by car availability.

Trips by mode must now be assigned to the different paths connecting the origin to the destination by that mode. Since each path implies a particular sequence of operators and transfers, trips are simultaneously assigned to operators, as well as to links of the network. This is carried out with another multinomial logit model, where the utility function is determined by the cost of each path. By applying vehicle occupancy rates, trips are transformed into vehicles by operator in each link of the network. In the case of public transport operators, exogenously given vehicles are assigned directly to the network; alternatively, the user can ask the model to determine the number of vehicles as a function of demand. Finally, the number of vehicles by operator are transformed into standard vehicles by applying appropriate rates.

In the final stage of the iterative process, travel speeds are reduced and waiting times are increased in every link for each operator as a function of demand/capacity ratios. Once travel times and waiting times have been adjusted, the calculation returns to the estimation of cost, because the new travel times will affect travel cost at path level, and consequently at all other levels. In turn, the new costs will affect trip generation, modal split and assignment, and even the location of activities in a future time period.

8.1.4 The evaluation procedure

For the simulation and evaluation of land use and/or transport policies, the model must be applied throughout the projection period at discrete time intervals for a base case scenario, where the policies are not included, and to an alternative scenario which includes an explicit definition of the policies. The differences in the results will represent the net effect of introducing the policies. At the end of the process an evaluation procedure can be applied, which compares the results of the base case and alternative scenarios, and estimates a number of socio-economic, fuel consumption and financial indicators.

In order to perform energy evaluation, the activity location model calculates the amount of fuel required for domestic heating as a function, among other variables, of floorspace density, insulating parameters and heat loss parameters. In the case of transport related energy consumption, the transport model estimates fuel consumption by operator as a function of speed. The results are read by the evaluation procedure which transforms consumption into standard energy units.

For the economic and financial evaluation it is necessary to provide the evaluation procedure with details such as the capital costs of the alternative scenario through time, discount rates, and shadow prices for some elements, such as for different types of fuel. With this data, the procedure will estimate indicators such as cost/benefit ratios, net present values and internal rate of return, considering economic or financial costs.

8.1.5 Operative structure of TRANUS

From an operative point of view, TRANUS consists of a number of computer programs linked to each other through data files. Figure 8.3

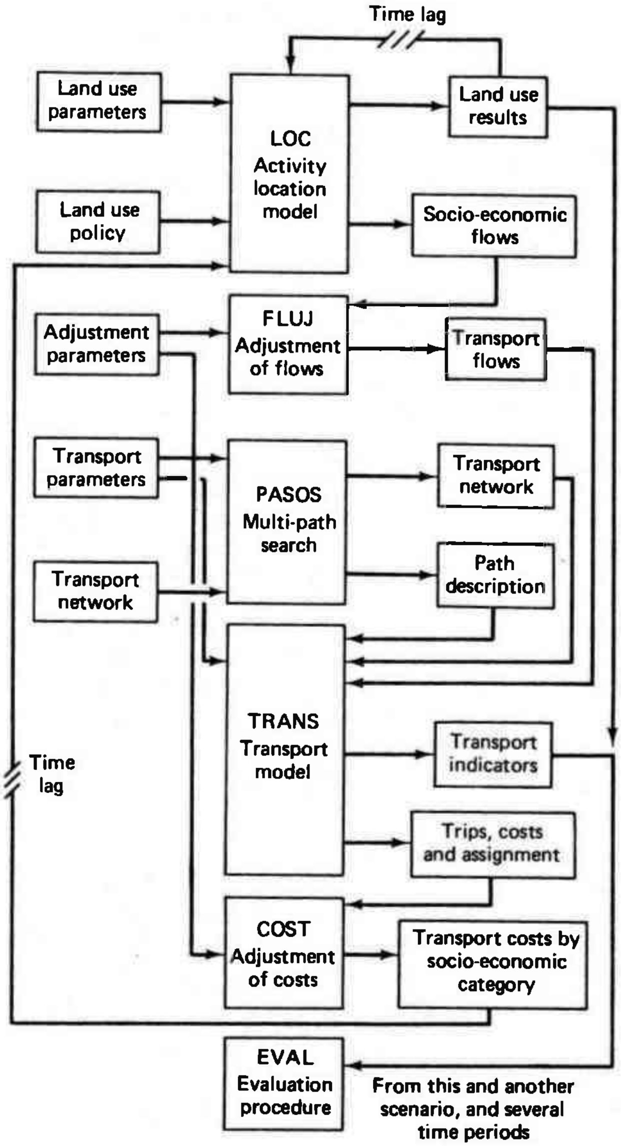


Figure 8.3. Operative structure of TRANUS

shows its main components. The main programs are:

LOC This program performs all the land use related calculations. It requires as input the location of activities in a previous time period, a description of current land use policy, transport costs for a previous time period, and a set of parameters. The program outputs the resulting location of activities for the current time period, together with floorspace estimates, land values, and locational utilities. It also produces a file containing the resulting matrices of functional flows.

FLUJ This program acts as an interface between the land use model and the transport model, transforming socio-economic categories into transport categories. These two types of categories may not be the same, and the analyst might want to change them in different ways, aggregating them, splitting them, or combining them. The program also allows for a number of additional matrix manipulations: matrices can be transposed, duplicated or combined. Some of the options are particularly useful to represent combined trip purposes, such as home–school–work, or to represent a proportion of empty return trucks. For 24-hour simulation, matrices are usually duplicated. FLUJ reads the matrices produced by the land use model and a parameter file where the optional transformations are specified, and produces a new file containing the resulting matrices by transport categories.

PASOS This program performs the multi-path search in the transport network. It will read a network file and a parameter file shared with the transport model. It will then search for the first *n* paths, connecting each origin to each destination by mode, storing the path descriptions in a set of output files.

TRANS This program contains the transport model. It will first read the network file, the path descriptions, and the functional flows by transport category produced by FLUJ. With this data, the program will perform all transport related calculations, i.e. trip generation, modal split, assignment and capacity restriction. Once supply–demand equilibrium has been reached in an iterative way, the program outputs the results of the assignment process, a file containing costs and trips matrices, and a file containing the evaluation indicators calculated during the simulation.

COST The purpose of this program is to act as a transport-to-land use interface. Thus, its function is the opposite of FLUJ. It will read the matrices of cost by transport category produced by the transport model, and will turn them back into socio-economic categories, following the same instructions given to FLUJ by the analyst, but in reverse order. It will also perform other functions, such as estimating costs for the first hierarchical level of zones when necessary, estimating intra-zonal costs, and making all resulting matrices symmetrical. The latter are stored in an output file.

EVAL This contains the evaluation procedure. Its purpose is to read the land use and transport outputs of two scenarios simulated for several time periods. With this and additional data from a parameter file, the program produces a set of comparative tables for the two scenarios through time. The comparative tables are complemented by a number of indicators that will help the analyst in policy evaluation. An accounting system, organised around users, operators and administrators, is compiled for the various time periods, calculating, among other things, present values of current costs and benefits on a yearly basis, present values of capital costs, consumers' surplus indicators, cost/benefit indicators and rates of return, together with energy evaluation.

8.2 Urban land use planning applications

8.2.1 Development plan of Curaçao

The purpose of this application was the design of an urban development plan for the island of Curaçao (1981 population, 147 388), and its main city, Willemstad. In 1976 a master plan had been designed which included broad proposals related to land use and the transport network (DROV, 1976). This master plan had to be updated, and the broad proposals had to be specified with more precision.

The TRANUS system (DROV, 1986) was calibrated for the base year 1981 to coincide with the 1981 census, which collected most of the required data, including a rudimentary form of origin–destination data and other transport variables. The census zones, which were related to a geo-coding system, were aggregated to 30 transport zones to cover the whole island. The zoning system and transport network are shown in figure 8.4.

Population was represented as households of three income groups, four types of land were considered, and employment was divided into four categories. For the transport simulation, trips to work for each income group were treated separately, as well as trips to services and trips to school. The network was coded with seven link types, which could be used by private or public modes. In turn, the public mode was divided into bus and minibus operators.

The simulated values of activity location, land values, peak hour origin–destination flows and traffic volumes were checked against existing data. Given the quality of data and the care with which it had been collected, it was possible to make a detailed calibration which resulted in very close real to simulated values.

Past trends in the growth of basic activities at the island level, together with a broad study of the national economy, were used as a basis for predicting future basic employment growth. Since at the time of this application economic prospects for the island were not considered good, two assumptions were adopted for the future, called 'trend' and 'low'. The latter particularly resulted in negative growth for most sectors.

From the point of view of land use policy, four alternative policies were devised: a trend or business as usual policy, a concentrated policy, an east-bound policy and a west-bound policy, depending on which direction the land use policy placed more emphasis. Land use policies were combined with the trend growth assumption, and the low growth assumption was tested only with the trend land use policy, resulting in a total of five scenarios in all. Each scenario included particular public housing projects and a common road building programme. Separate

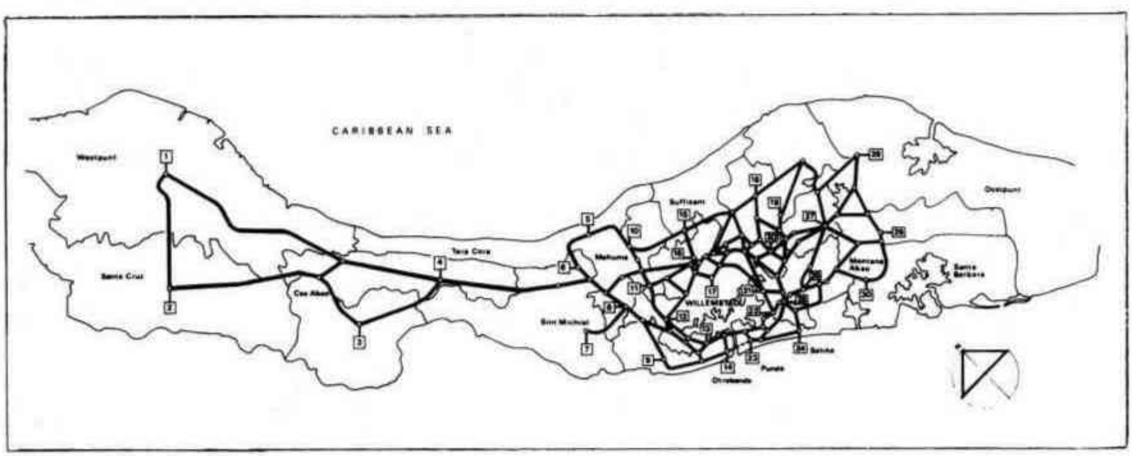


Figure 8.4. Zoning and transport network for the study of Curaçao

runs of the model were carried out from the base year 1981 to 1995 every five years. The first projection run was for the known year 1985, for which check data had been collected, making it possible to verify the dynamic behaviour of the model.

The results obtained proved useful in a number of ways. The main purpose of the simulation exercise was to define the emphasis of where development should go. The 1977 master plan had suggested that the west-bound scenario should be preferred. The results of the simulation, however, proved that extensions to the urban area in general had negative effects from many points of view, and that the existing urban area had sufficient capacity to absorb growth and expected changes in land use. Consequently, the land use policy that was finally adopted was characterised as 'relative concentration', together with a housing programme based on small in-filling projects instead of larger estates in the outskirts. The results of the model were also used to dimension the housing programme. On the transportation side, it was proved that the bus system had capacity problems, and that the increased revenues of a larger fleet would pay for the necessary capital costs. In fact, more buses were bought, which proved that the results were correct. The results from the assignment model proved that a major ring-road project was redundant, and was consequently eliminated from the investment programme.

The main conclusion from this application is that an integrated land use modelling package can be very useful for the design of an urban development plan, mainly because it provides consistency to the generally large number of policies that are part of the plan. For instance, if the tourist industry is assumed to play an increasingly important role in the economy of Curaçao, thus increasing the number of jobs in the centre of Willemstad, then a concentrated land use policy improves the work–home relationship (as well as the urban environment), and major road construction in the outskirts becomes unnecessary.

8.2.2 The master plan of La Victoria

The purpose of this application was very similar to that of Curaçao, i.e. the application of the TRANUS system to guide the design of an urban land use plan. The results are reported in MINDUR (1986).

The urban area of La Victoria in Venezuela (1981 population, 108 975) comprises three towns, and as shown in figure 8.5, is part of a larger surrounding region. Since it was known that there were strong interactions between La Victoria and the rest of the region, zoning was arranged in a hierarchical fashion: four regions were identified in the first level, of which La Victoria was one, in turn divided into zones 5 to 29.

Data was collected from a number of sources, and comprised four employment categories, three population categories, and three land use types. Private and public modes were distinguished in the transport system, the latter divided into buses and minibuses with different fare

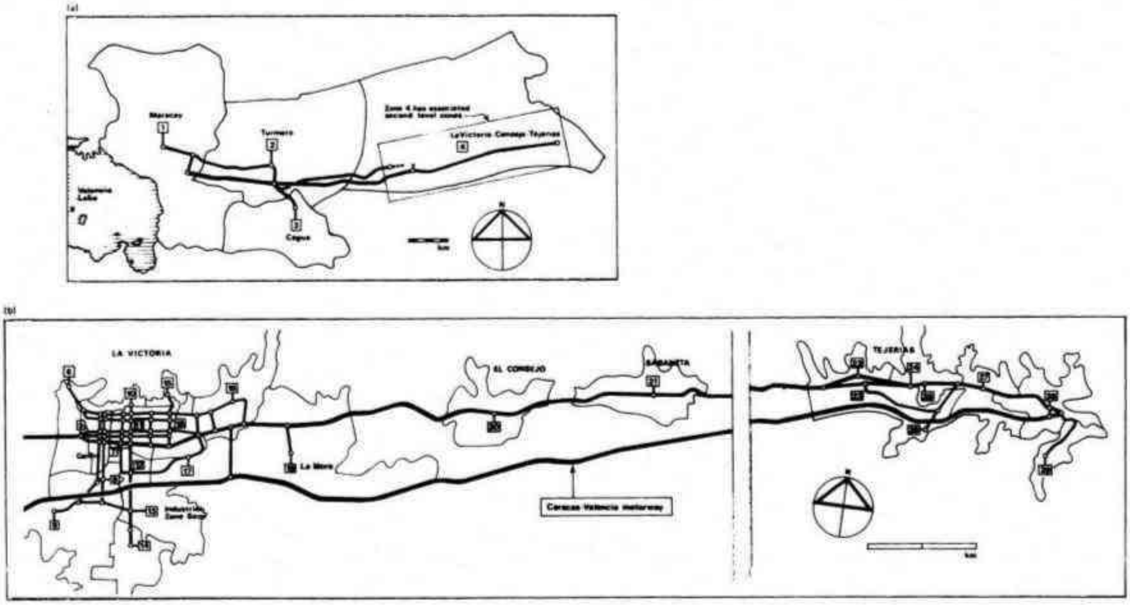


Figure 8.5. Hierarchical zoning and transport network for the study of La Victoria: (a) first level zones 1 to 4; (b) second level zones 5 to 29 (associated with first level zone 4)

structures. With this data, the model was calibrated for 1981 and 1985 in order to check its dynamic behaviour.

During the 1970s the study area went through a period of booming expansion, with an average of approximately 7% annual growth for basic industry, with similar rates for other sectors and population. However, from 1981 onwards, growth rates decreased substantially. As a result, two possible scenarios were devised: a high scenario, with an average annual rate of 3.5%, and a low scenario with a rate of 2.5%. Projections and simulations were carried out every five years from 1985 through 2005.

Instead of adopting several land use and transport scenarios, as in the case of Curaçao, a single scenario was simulated here, based on an iterative approach. Initially, a set of policies were devised and simulated with the model through the projection period. Then the results were examined and modified to produce a second set of runs. The process was repeated several times until the planning team was satisfied that the policies represented the best option. Each time, simulations were carried out for the high and low growth scenarios, in order to check the robustness of the proposals under different growth assumptions.

As a result, the model was not only useful for the process of policy design, but an integral part of it. Furthermore, in this application, as well as others not reported here, the model has been installed in the agency responsible for planning and is regularly used to increase the level of detail of the proposals, and to update them as new events emerge. For instance, soon after the analysis had been finished, a major railway project which seriously affects La Victoria came on the scene; the model was then used to analyse these effects and modify the proposals.

From a methodological point of view, the adoption of a hierarchical system of zones proved that La Victoria was seriously competing with other parts of the region for the attraction of new industry and housing developments, and that, as a result, was not able to maintain its relative importance. Towards year 2005, however, its relative importance would grow again, due to saturation of other areas. The hierarchical system was also useful to predict traffic flows along the main roads connecting La Victoria to its surrounding region, as well as significant through traffic. In fact, results suggest that through traffic will create serious congestion problems inside the urban areas. These conclusions would not have been possible if the hierarchical arrangement had not been adopted.

8.3 Urban transportation planning applications

8.3.1 Extension of the Caracas metro system

Figure 8.6 shows the metropolitan area of Caracas and surrounding areas (1981 population, 4.5 million). It is characterised by high density housing, a large concentration of government and tertiary employment, and a road network based on a relatively small number of urban motorways. These characteristics, combined with relatively high car-ownership rates and cheap oil prices, have congested the road system to a considerable extent. As a result, the city became a candidate for mass transit as early as 1966, when studies for the construction of a metro system began. Figure 8.6 also shows the proposed metro system with the sections that have already become operative.

The metro system has been very successful in providing a high quality

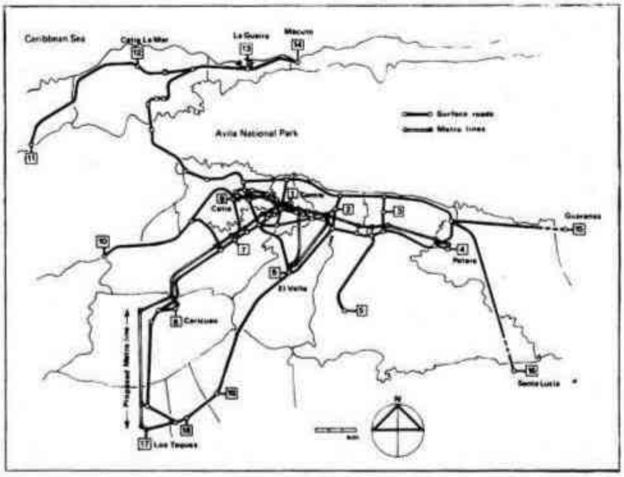


Figure 8.6. Zoning and transport network for the study of the extension of the metro line in Caracas

service, but can do very little for the increasing number of people living in the so-called expansion areas, valleys which are located 20 to 50 kilometres away from the main urban area, separated by mountains. In order to contribute to the solution of this problem, the Caracas Metro Company decided to study the feasibility of building a line connecting Caracas to Los Teques, one of the expansion areas, as shown in figure 8.6.

For the estimation of the demand, the TRANUS system was applied to the whole of the urban region of figure 8.6. The inclusion of such a large area in the analysis was considered necessary, because expansion areas compete for development; hence improving the accessibility to one of them affects all others. Results of this application are reported in CAMETRO (1987).

Since the study had to be carried out in a very short period of time, the area was divided into a small number of large zones, also indicated in figure 8.6. The transport network, however, had a certain complexity, due to the fact that it had to adequately represent urban and suburban buses, minibuses, the existing and future metro lines, urban and regional motorways, main roads, metro stations, pedestrian links, and so on.

Calibration of the model was carried out for 1981 and 1985, and a single growth scenario was assumed for the future, again in five-year periods to 2005. Within this limited scope, two future scenarios were simulated and evaluated: with and without the proposed extension of the metro system. The rest of the elements, such as land use policy, public housing, the rest of the metro expansion programme and other transport related policies were deliberately kept unchanged for both scenarios, to provide a *ceteris paribus* character to the analysis.

The results went well beyond its main purpose; mainly the dimensioning of the proposed system and whether its construction should be recommended at all. The study did show that the extension was beneficial from many points of view, both to users and to the operation of the metro system in general, although it raised doubts abouts its cost-effectiveness due to high capital costs. The results were used to determine the size and frequency of trains, the capacity of the stations, and fare structure of the service. But the study also provided insights about the functioning of the whole metro system in the future. For instance, it showed the points in the system where alarmingly high demand levels will be reaching capacity, both along particular rail links as well as some of the pedestrian connections. Additions to the current programme are being devised to help solve some of the conflicts. The analysis also served to highlight the potential of the Los Teques area to absorb urban development, and the closeness with which such potential is related to accessibility. The proposed extension, it was concluded, will not only serve a purely transportation purpose, but will allow for the incorporation of a large residential area, reducing the cost of living to a considerable portion of the population.

8.3.2 A monorail system for the south-east of Caracas

The south-east of Caracas is a large area with great potential for medium to high income residential development. In contrast to the area of Los Teques described in the previous section, this is not strictly an expansion area, but, due to its closeness and lack of important physical boundaries, must be considered an integral part of Caracas.

Access to the centre of town, however, is limited to two major roads, which are clearly insufficient for an increasing demand with high car-ownership. Since new roads cannot be introduced and the current ones cannot be expanded, local authorities decided to 'freeze' building permits in the area, igniting a conflict between current and potential dwellers, authorities and private developers. In order to solve this problem, a private group proposed the construction of a monorail system that would connect the south-east with the centre, acting as a metro feeder at the same time.

In order to dimension the monorail system, the stations and a complementary bus-feeder system, the TRANUS model was applied to exactly the same area shown in figure 8.6 above, but since the complexity of the area is much larger than that of Los Teques, a second hierarchy of zones was introduced, as shown in figure 8.7. The modelling exercise,

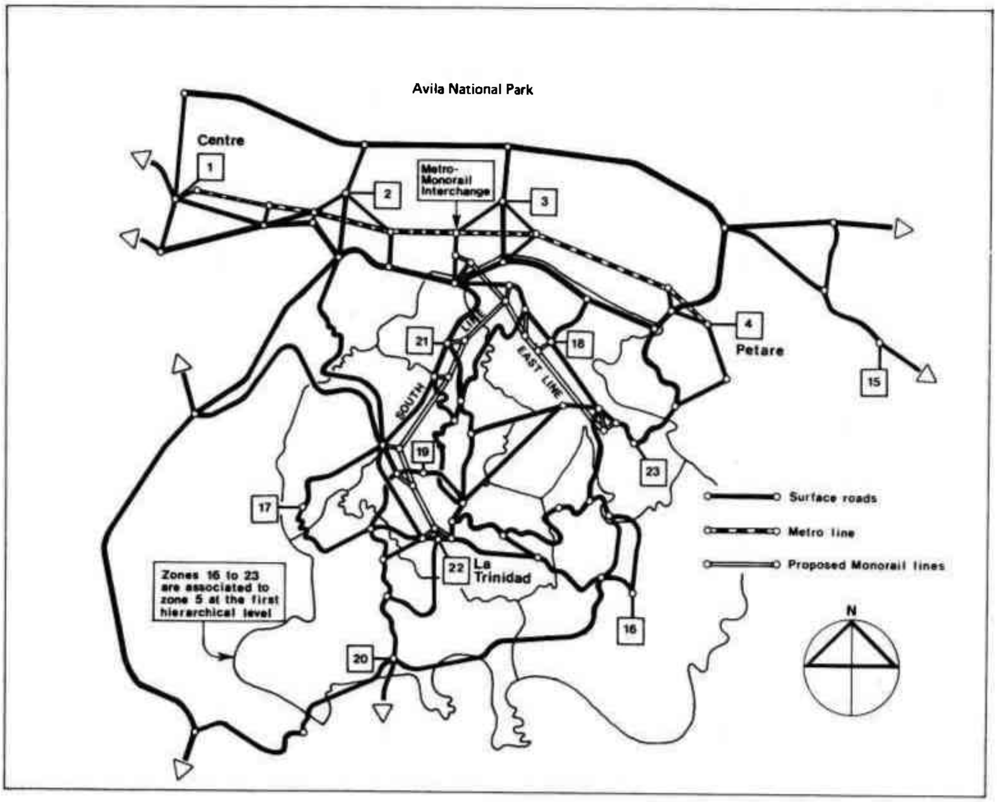


Figure 8.7. Zoning and transport network for the evaluation of the monorail system for the south-east of Caracas

reported in Aerorriel (1987), also had the purpose of evaluating the effects of the proposed system on the population and on the rest of the transportation network.

Two sets of scenarios were defined for short-term projections. A first set included three broad alternatives for the study area: a do-nothing scenario, a scenario based on a bus system on exclusive lanes, and a monorail based scenario. A second set of scenarios was simulated in order to determine the ideal set of routes for the system, and the fare structure that optimised its financial feasibility.

Results from the first set showed that the do-nothing option led to medium-term chaos, while the exclusive bus lane option led to a short-term chaos. In the first case, results showed that authorities were right in implementing the 'freezing' policy for the area. The exclusive bus lane system would provoke immediate chaos because the only possibility was to extract capacity from the existing and congested roads to the exclusive lanes. It was interesting to see that the model predicted some benefits for the bus-riders of the area, but inflicted great pain to car-owners and even bus-riders in the rest of the city. Results from the monorail scenario, by contrast, showed a general improvement throughout the city, and mixed well with the metro system. They also showed that the monorail was not able to perform its functions, unless complemented with a minibus feeder service.

The second set of scenarios was devised to help in the design of the routes and stations, as well as the feeder system. Since the proposed system is to be implemented by the private sector, design considerations and demand analysis were closely related to financial feasibility. A large number of tests were carried out, varying fare structures, location of the stations and route layouts, until a particular design was achieved which produced estimates of income enough to pay for operation and capital costs. In order to achieve this, the original design of the monorail evolved considerably through the modelling exercise.

8.4 Regional transportation applications

8.4.1 The Caracas–La Guaira motorway

Caracas is linked to the coast through a double-lane motorway. Along the coast there is a thin strip of land squeezed between the mountain and the sea. The area is densely occupied and performs several functions. Venezuela's main port and airport are located here, and there are many recreational facilities together with weekend housing. Due to its closeness to Caracas, it also acts as an expansion area, similar to Los Teques described above, housing a population of 400 000. This forces many people to commute, and as a result the number of trips along the motorway has increased steadily through time. During weekends recreational trip-makers must queue for hours.

The Caracas–La Guaira motorway also has many physical problems, particularly landslides, which increase its maintenance costs, partly compensated by a toll system. In order to solve some of these problems, several alternative projects have been devised, from extensions to parallel roads and railways. Recently, the government has been studying the possibility of privatising maintenance, but in order to negotiate with future concessionaires, required a full study of future demand.

To this end, the TRANUS model was applied to a zoning system very similar to the one used in the study of Los Teques, except for the coastline area which was disaggregated into several zones. Results of this study are reported in MTC (1986). As in the previously described application, calibration was carried out for 1981 and 1986.

Projections were made every five years through to 2005, considering several scenarios. This time scenarios were defined as future events that would affect the management of the motorway. A first trend scenario assumed that there were no projects around the existing motorway; a second scenario assumed the introduction of a second motorway, roughly parallel to the existing one; a third scenario assumed the introduction of a passenger railway system, also parallel to the existing motorway; a fourth scenario included both the motorway and the railway.

The first interesting result from the model was that in the base year 1986, the toll system had to collect at least twice as much as what was being collected, a similar conclusion to the one arrived at by a congressional committee set up to investigate corruption in toll stations. The model showed that the toll stations should collect enough to cover maintenance costs of the motorway, of an old and winding road that also connects Caracas to La Guaira, and of parts of the road system in both ends. This sum, according to the model estimates, would increase in the future but much more slowly than expected by previous studies. The reasons for this slow growth are two-fold: the capacity of the existing motorway restricts demand, and the capacity of both La Guaira and Caracas to house new development is also limited.

Results from the other scenarios showed an increase in demand as the first of these restrictions was removed, but growth was still limited by the second one. The parallel motorway was, in general, more successful than the railway system, because it attracted traffic of all types, while the train service only included passengers. Also, since both solutions must have a relatively steep gradient, buses can achieve higher speeds than trains.

Paradoxically, from the concessionaire's point of view, the do-nothing trend scenario was the best. The parallel motorway scenario doubled his maintenance costs and increased revenues by only 30% because of land use restrictions. The railway kept his maintenance costs constant but reduced revenues. Naturally the addition of both motorway and train was the worst scenario for the concessionaire. Some of these results are represented diagrammatically in figure 8.8.

Apart from the future estimates that were the main purpose of the study, results from the model highlighted two other main conclusions: (a) land use and environmental restrictions in the La Guaira area

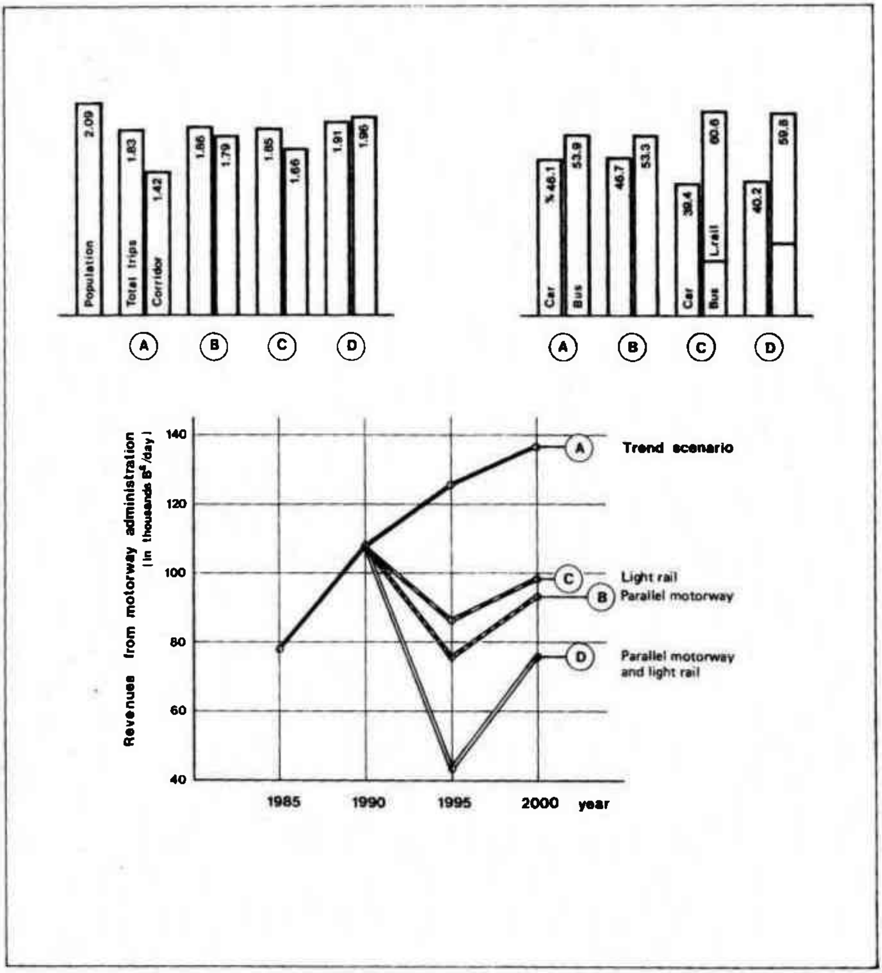


Figure 8.8. Some results of the study of the Caracas–La Guaira motorway. Left: Annual growth rates from 1985 to 2000 of population, total trips, and trips along the Caracas–La Guaira corridor. The trend scenario (A) is the most congested particularly in the corridor, hence the smaller number of trips. When the parallel motorway is introduced (B) the number of trips along the corridor increases substantially. The light rail scenario (C) generates fewer trips in the corridor. The combined motorway and light rail scenario (D) accommodates the most trips, but still does not match population growth. Right: Modal split along the Caracas–La Guaira motorway – year 2000. When the parallel motorway (B) is introduced, there is a small shift to the private mode in comparison with the trend scenario (A). When light rail is introduced (C), use of public transport increases substantially, although use of buses declines. Bottom: The difference between maintenance costs and income from tolls. Revenues are largest in the trend scenario (A). When the parallel motorway is introduced (B), maintenance costs almost double, but income does not increase accordingly. When light rail is introduced (C), maintenance costs remain unchanged, but income drops. When both projects are introduced (D), the motorway administrator must absorb substantially increased maintenance costs with reduced income

question the convenience of building new transport facilities to the area, and (b) of these, the railway is the least convenient because of its limitations under steep conditions, and high capital costs. In fact, the government announced recently a programme to improve the existing motorway and permanently dropped the railway project which had been completed from the engineering point of view, diverting the funds to other more convenient projects. From a methodological point of view, the study demonstrated that it is impossible to make even medium-term projections of a road like the Caracas–La Guaira if the model does not integrate land use and transportation.

8.4.2 The Central Railway system of Venezuela

Venezuela has only one small railway line connecting the city of Barquisimeto with Puerto Cabello. For some time, the railway authority (FERROCAR) has been studying an ambitious programme to build a complete network for the country, and the study of the first line, connecting Caracas to Valencia and Puerto Cabello, is well advanced. In fact, it was FERROCAR who carried out and later dropped the Caracas–La Guaira project referred to in the previous section.

A railway link between Caracas, Valencia and Puerto Cabello is of strategic importance to the economy of the whole country. The TRANUS system was applied, then, to the country divided into regions, and a second hierarchical level was defined for the two regions directly affected by the railway project. This is shown in figure 8.9. The results of the study, which was carried out together with the firm TRANSPLAN, are reported in FERROCAR (1987). The purpose of the study was two-fold: to estimate future demand for the layout and dimensioning of the track, trains, stations and services, and to assess the spatial and socio-economic effects of the system.

This is by far the most complex application of the TRANUS system so far. The model was asked to simulate the location of population and employment throughout the country distinguishing economic sectors and income groups. External zones were defined to represent imports and exports. It had to represent the movement of commodities of different types (light and heavy bulk and general cargo) using light and heavy trucks, and railway. It also had to represent population by income groups performing trips to work and to services using cars, regional and surburban buses, and regional and surburban passenger trains. In addition, access to cities involved urban buses and trains (metro links).

To calibrate such a complex system, special origin–destination roadside interviews were carried out at strategic points in the network. All lorries were stopped and information about the particular commodity being carried and the origin and destination zone and sector was collected. All bus drivers were similarly interrogated, complemented by data collected at bus stations. A sample of cars was stopped to find out origin, destination and purpose of trips.

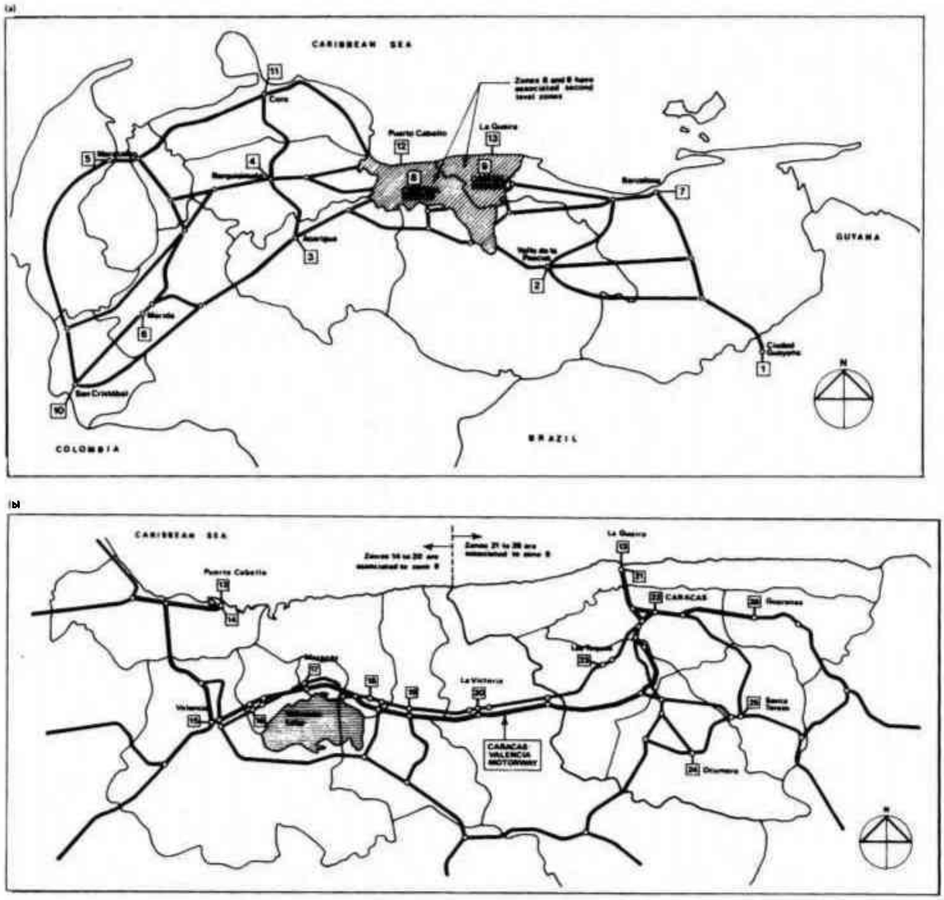


Figure 8.9. Zoning and transport network for the study of the Caracas–Valencia–Puerto Cabello railway: (a) first level zones 1 to 13; (b) second level zones 14 to 26 (associated with first level zones 8 and 9)

After an extensive calibration effort, projections were carried out through to 2010. First, two scenarios were tested: with and without the proposed railway. Results showed that demand for both passengers and commodities would grow steadily through time, and that the railway would have a definite and substantial impact on the location of activities. The land use model showed the emergence of a large linear city developing from the south of the Central Region (the Tuy Valley) right through to Valencia. From the transportation point of view, results showed that if the railway is not built, the main Caracas–Valencia motorway will face medium-term congestion, deteriorating its level of service alarmingly. Demand on the railway system went as high as 250 thousand passengers on line per day.

Having established the convenience of building the railway, a second set of runs was carried out with design purposes, in order to determine the purpose, size and location of stations, and the frequency and capacity of regional and suburban services. Tests were also carried out to estimate demand sensitivity to fares, and the competition between trains and buses and lorries. Currently the engineering project and the operational plan are being developed; for this purpose the model is consulted in a permanent way. The model has been installed in FERROCAR to evaluate other parts of the national railway system.

From a methodological point of view, the main conclusion is that a very complex application like the one described is perfectly possible at reasonable costs. The package proved to be light enough to enable a large number of tests, in spite of the fact that personal microcomputers are being used. From a theoretical point of view, this application shows that the set of nested mulitnomial logit models can be successfully applied on a nationwide scale. The multi-path search algorithm allowing transfers from regional buses to trains to metro and urban buses performed well, producing realistic results. Since the network was perfectly symmetrical and 24 hour demand was also assumed to be symmetrical, symmetry was expected and obtained from the assignment module, proving that the algorithm is independent of the topology of the network. Traffic counts showed that directional variations in traffic are very small (less than 5%), and due to the randomness of their distribution, differences are attributable to collection errors rather than to an asymmetrical behaviour.

8.4.3 Evaluation of five motorway projects

The same system described above for the railway system was later applied for the evaluation of five motorway projects which were part of a bid by the Ministry of Transport of Venezuela to the Interamerican Development Bank (BID). The results obtained are reported in MTC (1987). In order for the bid to be successful, the bank must be satisfied that each project produces an internal rate of return above 12%.

Traditionally, road building projects are evaluated by assuming a growth rate for traffic for a period of fifteeen to twenty years. Operating costs of vehicles are then estimated on a with and without project basis. The resulting savings are then multiplied by the expected traffic obtaining net present values, which are compared to the net present value of capital costs to estimate the internal rate of return.

However, each one of the five projects is of strategic importance to the country, and hence the simple assumption of growth rates for traffic is unsatisfactory. The TRANUS system was used instead to obtain traffic projections, and since the package includes an extensive evaluation procedure, the resulting rates of return were compared to the ones obtained in the traditional way. Projections were carried out through to year 2005 according to seven scenarios: a scenario with none of the projects, one scenario for each project, and a scenario with all projects simultaneously.

For each project three traffic projections were obtained: without the project (in the existing old road), with the project and with all projects. The results showed, in the first place, that the projects were not independent from each other. In one particular case, because of the way traffic was distributed system-wide, the project showed, when evaluated in isolation, a rate of return of 7%, that is, below the minimum; the same project, evaluated as part of the set of five, showed a rate of return of 43%, well above the minimum. The main conclusion here is that projects of this kind cannot be evaluated in isolation, but only as part of a broader system.

It was also interesting to compare the results from TRANUS with the ones obtained by the traditional method. Savings in operating costs are not the only variable to be included in the evaluation and there are many other positive or negative aspects that must be taken into consideration. In another case, for example, the proposed road concentrated traffic along an existing motorway which is currently operating close to capacity; benefits perceived in the project itself were compensated negatively by the congestion it created in other parts of the network.

In general, a regional or urban transportation system is far too complex with many tight relationships with other transport elements and the activity system. It is surprising, then, that large amounts of money should be decided on the basis of extremely simple and short-sighted methodologies.

8.5 Land use and transport energy consumption

8.5.1 Alternative scenarios for energy consumption in an urban region

This research has been carried out by the author with P. A. Rickaby at the Centre for Configurational Studies, The Open University, UK, and is reported in Rickaby and de la Barra (1987). The starting point of the study was the relationship between the spatial organisation of society and its use of energy. The conventional wisdom holds that compact settlements are fuel conserving, and that if fuel is to be saved, then the spread of surburbia must be halted and new development directed into existing urban areas. Such policies are intended to reduce the separation of homes and work-places and encourage a switch from private cars to the more fuel efficient public modes of transport.

Taken to the extreme, the most fuel efficient settlement is one in which nobody travels or does anything. Therefore what is required is not absolute fuel conservation but relative fuel efficiency. In other words, the ideal settlement is the one that saves the maximum amount of fuel with existing technology, and at the same time allows for maximum mobility, or as it turned out in this study, can be achieved at the least social cost.

The purpose of the study was two-fold: (a) to develop a methodology that could assess costs and benefits in terms of energy, and (b) to explore a set of realistic options for the future development of a typical urban region and assess them from the point of view of energy use.

The TRANUS system estimates the use of two types of energy: within place and between place energy, i.e. in dwellings and transport. Domestic space heating is calculated as a function of the floor area consumed by each activity type in each zone, the overall conversion efficiency of the heating systems in the dwellings and a *retrofit parameter* which allows for changes in the market penetration of central heating systems or changes in the level of insulation. Fuel use by vehicles of different types is calculated as a function of average speed in each link of the network after capacity restriction.

Data was collected for a sample of twenty British cities with populations of between 50 000 and 150 000 persons. Data from these cities were combined with national statistics to produce an archetypal configuration and an associated 52-zone distribution of population, employment and transport system, in order to free the exercise from local irregularities. The model was calibrated against this idealised data-set which is diagrammatically represented in figure 8.10 as pattern 0. Data for the calibration of the domestic heating function were obtained from studies of the British housing stock by Penz (1983) and Leach and Pellew (1982). Vehicle consumption was obtained for typical British vehicles of several types.

From this starting point, five alternative patterns were devised to represent realistic possibilities for the development of city regions and to compare their properties in respect of accessibility and energy use. The modifications of the existing configuration were all based on published speculations about the way in which settlement patterns might be altered in order to save fuel in buildings, or transport, or both. They include the concentration of new development into the central city (pattern 1), into ribbon developments (pattern 2), and into satellite towns (pattern 3), and the dispersal of new development into both linear and nucleated configurations (patterns 4 and 5). All of the modified patterns included the same network of major roads, the same central area, and the same total city population as derived from the existing configuration. In each case the modifications are the assumed result of strategic planning policies operating over a period of 25 years.

In addition, the comparative evaluation of the patterns was made in the context of three energy scenarios. The scenarios characterise the range of realistic possibilities for economic and energy supply conditions in Britain on a 25 year horizon. They range from a high energy business-as-usual scenario, through a technical-fix scenario in which economic growth is decoupled from growth in energy use, to a very-low-energy scenario combining intensive fuel conservation with low economic growth and with changes in lifestyle and standards of living. Figure 8.11 shows the primary fuel demand in each of the three scenarios.

In order to compare the patterns, a standard unit was devised relating benefits with energy: utils/gigajoule saved (U/G). Utils was the name given to the indicator of average utility which combined land use and transport benefits (from equation (4.11)). The results obtained from the model, shown in figure 8.12, indicate that less fuel is used in each of the five modified patterns than in the original, but in every case there is a

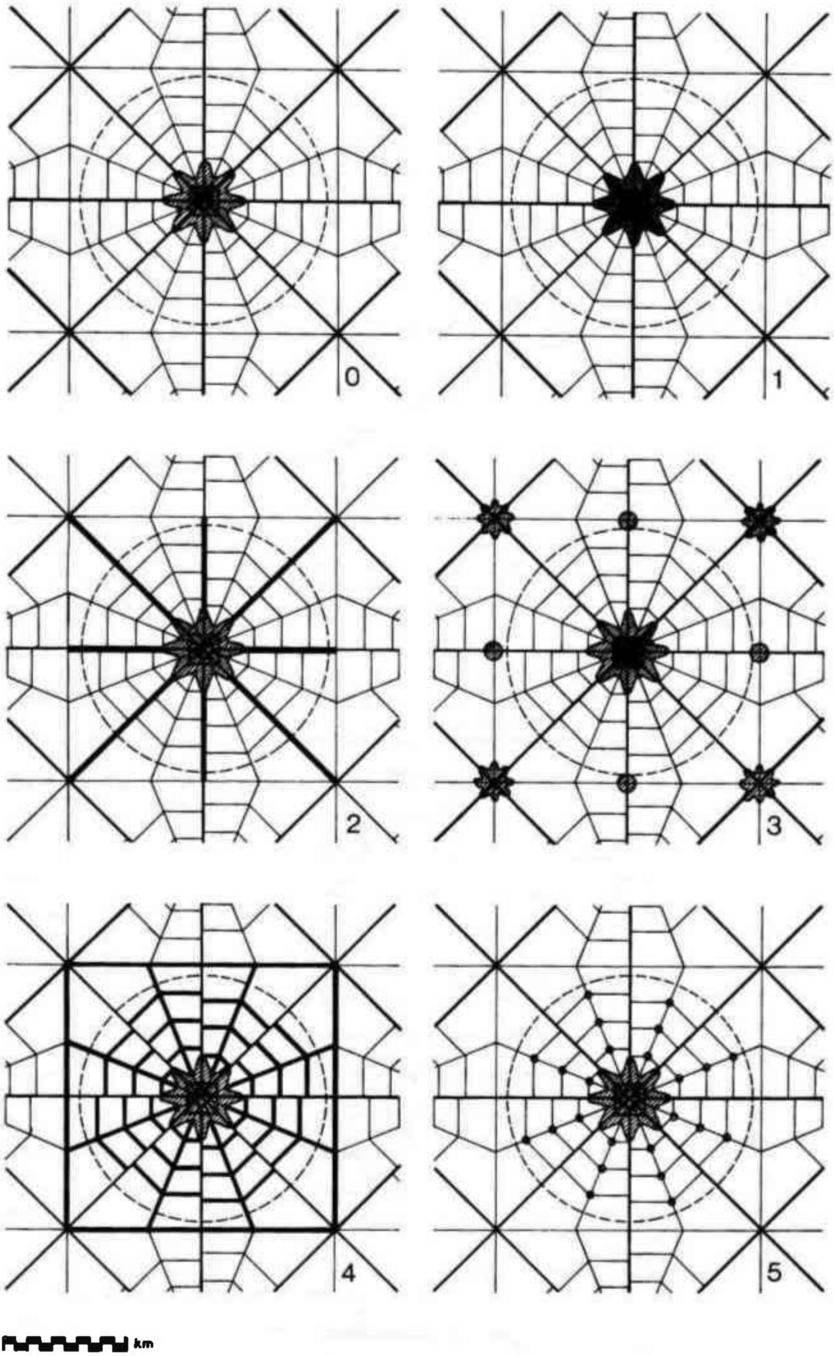


Figure 8.10. Six development patterns for energy evaluation

loss of benefits. The greatest saving of fuel, particularly from transport, is made in pattern 1, which is the result of a sustained policy of urban containment, increasing overall density by about a third and reducing trip lengths considerably.

Figure 8.11. Primary fuel demand for the three energy scenarios in the energy evaluation study

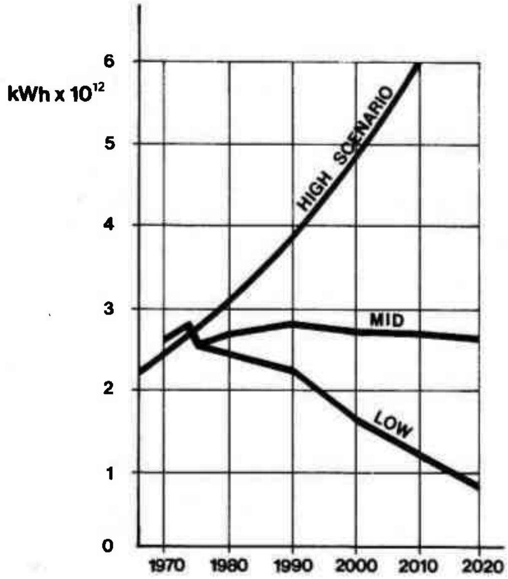
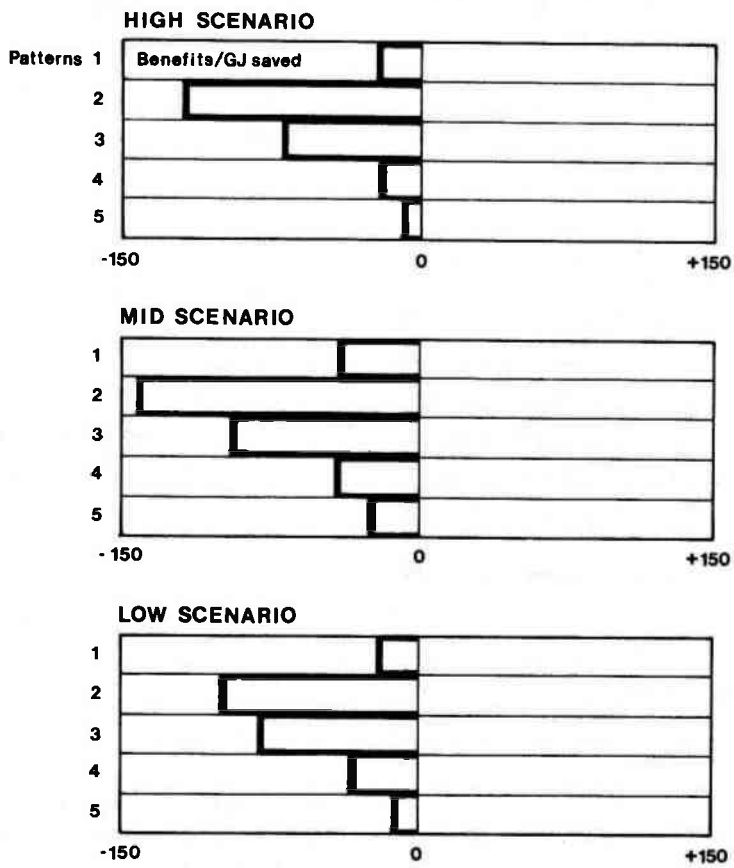


Figure 8.12. Overall energy evaluation summary for each combination of scenario and pattern



The next largest fuel savings are made in pattern 3, based on satellite towns at some distance from the central city. However, these savings are almost exclusively in the domestic sector, compensated by longer trips. The fuel savings in this pattern are the most costly, varying from 70 to 93 U/G according to scenario. The least costly fuel savings are made in pattern 5, in which population is located in villages, varying from 10 to 25 U/G. In the transport sector there are substantial positive benefits associated with fuel savings: 55 U/G in the high scenario, 234 U/G in the mid scenario, and 200 U/G in the low scenario. Patterns 2 and 4 are linear configurations that produce savings in domestic fuel against an increase in transport fuel due to the concentration of traffic along main roads.

It is interesting to consider the results from the point of view of the energy scenarios. In a business-as-usual environment, the existing settlement pattern appears to be the most appropriate. In a technical-fix environment aimed at fuel conservation without changes in lifestyle, patterns 1 or 5 would be the most appropriate targets for planning policy. In a very low economic growth scenario with intensive fuel conservation, patterns 1 and 5 again provide the most appropriate policy targets, but the high urban densities associated with pattern 1 may restrict the exploitation of solar energy for heating, which is an important component of this scenario.

From a general methodological point of view, this application shows that integrated land use and transport modelling can be very useful to assess policy implications on the consumption of energy. The integrated approach is essential because within place and between place energy consumption are closely related. In the application reported here, the modelling system has been applied to a hypothetical town, but it can be equally applied to any real case study. Furthermore, the complexity of the calculations involved suggests that integrated modelling is, perhaps, the only method currently available for this kind of analysis. Hence, energy evaluation represents one of the most promising areas of application for integrated land use and transport modelling for cities of Europe and the USA.

Appendix: computer programs

Throughout this book, reference has been made to a number of computer programs that are intended to help the understanding of particular topics. They can be used by the interested reader for his own benefit, or they can be used for computer aided teaching.

All programs have been written for the MS–DOS operating system and require modest resources, so that they should work on many popular personal computers. There are English and Spanish versions. In some cases, there are programs that require graphic screens, but there is always an alternative version without them. Each program is accompanied with a document file to explain its use, and one or more example data-files.

Users interested in acquiring these programs should post two DSDD $5\frac{1}{4}''$ diskettes with a covering letter explaining the intended use (for the purpose of keeping a record) to the following address:

T. de la Barra
Apartado Postal 47709
Los Chaguaramos
Caracas 1041-A
Venezuela

What follows is a short description of each of the programs referred to in the book. All programs have been written by the author. In many cases they are representations of models proposed or developed by others, and they have been written on the basis of published material. Hence, there might be conceptual differences between these versions and the originals, not always involuntary.

DIAL

This program contains a simplified version of Dial's assignment algorithm, with a fixed number of trips from a given pair of origin and destination nodes.

ECHE

This program contains a version of Echenique's (1969) model of floor-space and activity location. It is similar to the program LOWRY, and in fact can use the same data-files, to facilitate the comparison of the results.

INPRO

This program contains the single-region input–output model. Solution is achieved through the iterative method proposed in figure 5.5. Intermediate outputs are shown to allow for a better understanding of the model.

LOWRY

This program contains an entropy maximising version of Lowry's (1964) original model. It has been written so that it can share the same data-files with program ECHE to facilitate the comparison of results.

REGINP

This program contains the multi-region input–output model as described in section 5.4. By setting the parameters in appropriate ways, the program can include elasticities and prices. Intermediate results can be inspected optionally.

RIVER

This program has been specifically designed for the example in figure 4.5. The user can run the same example or can modify data to experiment with variations.

UTIL

The purpose of this program is to calculate a number of numerical examples related to discrete choice and random utility theory. One or several options can be calculated and compared, as well as one or several decision-making groups.

VONTH

This program contains an extended version of Von Thünen's original model, along the lines explained in chapter 2. By fixing the parameters in appropriate ways, the bare original model can be run, then demand elasticities can be introduced, and finally land consumption elasticity. If the user has a graphic screen, a version of this program will draw the resulting diagrams.

References

Aerorriel (1987) *Proposicion de un Plan de Transporte para el Sur-Este de Caracas*. Aerorriel SA

Alonso, W. (1964) *Location and Land Use*. Cambridge, MA: Harvard University Press

Anas, A. (1982) *Residential Location Markets and Urban Transportation. Economic Theory, Econometrics, and Policy Analysis with Discrete Choice Models*. New York: Academic Press

Angel, S. and Hyman, G. M. (1971) *Urban Spatial Interaction*. CES WP 69. London: Centre for Environmental Studies

Barto, A. G. (1978) Discrete and continuous models. *International Journal of General Systems*, **4**, 307–32

Batty, M. (1975) In defense of land use modelling. *The Planner: Journal of the Royal Town Planning Institute*, **61**, 184–7

(1976) *Urban Modelling: Algorithms, Calibrations, Predictions*. Cambridge: Cambridge University Press

Baxter, R. (1976) *Computer and Statistical Techniques for Planners*. London: Methuen and Co.

Baxter, R. and Williams, I. (1973) *The Third Stage in Disaggregating the Residential Sub-Model*. WP 66. Cambridge: Land Use and Built Form Studies

(1975) An automatically calibrated urban model. *Environment and Planning A*, **7**, 3–20

Blunden, W. R. (1971) *The Land-Use Transport System*. London: Pergamon Press

Broadbent, T. A. (1973) *Activity Analysis of Spatial Allocation Models*. CES WP 82. London: Centre for Environmental Studies

Buchanan, C. (1963) *Traffic in Towns*. London: HMSO

Bunge, M. (1973) *Method, Model and Matter*. Dordrecht, Holland: D. Reidel Publishing Co.

CAMETRO (1987) *Demanda de Transporte Linea Metro Las Adjuntas–Los Teques*. Caracas: Compania Anonima Metro de Caracas

Campbell, D. T. (1969) Reform as experiments. *American Psychology*, **24**, 409–29

Christaller, W. (1933) *Die Zentralen Orte in Suddeutschland, Jena*. English translation by C. W. Baskin (1966): *Central Places in Southern Germany*. Englewoods Cliffs, NJ: Prentice Hall

Clark, C. (1951) Urban population densities. *Journal of the Royal Statistical Society, Series A*. **114**, 490–6

Cochrane, R. A. (1975) A possible economic basis for the gravity model. *Journal of the Royal Statistical Society, Series A*, **114**, 490–6

Converse, P. (1949) New laws of retail gravitation. *Journal of Marketing*, **14**, 279–304

Daganzo, C. F. (1980) Optimal sampling strategies for statistical models with discrete dependent variables. *Transportation Science*, **14**, 324–45

Dale, H. M. (1977) Trip generation: what should we be modelling? In *Determinants of Travel Choice*, D. A. Henscher and Q. Dalvi. London: Saxon House

de la Barra, T. (1980) *Proposicion de un marco teorico, unificado y consistente para la simulacion y evaluacion de planes de uso del suelo y transporte*. PT9, Instituto de Urbanismo, Universidad Central de Venezuela, Caracas

de la Barra, T. and Perez, B. (1986) Assymetry in some common assignment algorithms: the dispersion factor solution. *Environment and Planning B*, 13, 293–304

de la Barra, T., Echenique, M., Guendelman, J. and Quintana, M. (1974) *An Urban-Regional Model*. WP 78. Cambridge: Land Use and Built Form Studies

de la Barra, T., Perez, B. and Vera, N. (1983) *Estructuracion jerarquica de Modelos de Uso del Suelo y Transporte*. Instituto de Urbanismo, Universidad Central de Venezuela, Caracas

(1984) TRANUS-J: Putting Large Models into Small Computers. *Environment and Planning B*, 11, 87–101

Dial, R. B. (1971) A probabilistic multi-path traffic assignment model that obviates path enumeration. *Transportation Research*, 5, 43–111

Dijkstra, E. W. (1959) *A note on two problems in connexion with graphs. Numerische Mathematik*, 1, 269–71

Domencich, T. and McFadden, D. (1975) *Urban Travel Demand: a Behavioural Analysis*. Amsterdam: North-Holland

DROV (1976) *Master Plan for the Island*. Curaçao: Department of urban and Regional Development Planning and Housing (DROV)

(1986) *Application of a Land Use-Transport Model to Curqcao*. Curaçao: Department of Urban and Regional Development Planning and Housing (DROV)

Echenique, M. (1968) *Urban Systems: Towards and Explorative Model*. WP 7. Cambridge: Land Use and Built Form Studies

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. (1973) *A Disaggregated Model of Urban Spatial Structure: Caracas*. WP 8. Cambridge: Land use and Built Form Studies

Echenique, M., Geraldes, P. and Williams, I. (1977) *Desarrollo de un modelo de uso del suelo y transportes*. Informe No. 2 para la Corporacion Administrativa del Gran Bilbao. Cambridge: Applied Research of Cambridge Ltd

FERROCAR (1987) *Estudio de Ruta y Demanda del Ferrocarril Caracas–Tuy Medio–Valencia–Puerto Cabello*. Caracas: Instituto Autonomo de Ferrocarriles del Estado

Garin, G. A. (1966) A matrix formulation of the Lowry model for intrametropolitan activity location. *Journal of the American Institute of Planners*, 32, 361–4

Goodall, B. (1972) *The Economics of Urban Areas*. London: Pergamon Press

Hansen, W. G. (1959) How accessibility shapes land use. *Journal of the American Institute of Planners*, **25**

Herbert, J. and Stevens, B. H. (1960) A model for the distribution of residential activity in urban areas, *Journal of Regional Science*, **2**, 21–36

Hoyt, H. (1939) *The Structure and Growth of Residential Neighbourhoods in American Cities*. Washington DC: Federal Housing Administration

Huff, D. L. (1962) A note on the limitations of the intra-urban gravity model. *Land Economics*, **38**, 64–6

(1963) A probabilistic analysis of shopping centre trade areas. *Land Economics*, **39**, 81–90

Isard, W. (1956) *Location and Space-Economy*. Cambridge, MA: MIT Press
(1960) *Methods of Regional Analysis*. Cambridge, MA: MIT Press
Kanafani, A. (1983) *Transporation Demand Analysis*. McGraw-Hill
Kirwan, R. M. and Martin, D. B. (1970) *The Economic Basis for Models of the Housing Market*. CES WP 62. London: Centre for Environmental Studies
Kuhn, T. S. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press
Lakshmanan, T. R. and Hansen, W. G. (1965) A retail market potential model. *Journal of the American Institute of Planners*, 31, 134–43
Leach, G. and Pellew, S. (1982) *Energy Conservation in Housing*. London: International Institute for the Environment and Development/Science Reviews
Le Blanc, L. J., Morlok, E. K. and Pierskalla, W. P. (1975) An efficient approach to solving the road network equilibrium traffic assignment problem. *Transportation Research*, 9, 309–18
Lee, C. (1973) *Models in Planning*. London: Pergamon Press
Leontief, W. W. (1941) *The Structure of the American Economy 1919–1939*. 2nd ed. 1951. New York: Oxford University Press
Leontief, W. W. and Strout, A. (1963) Multi-regional input–output analysis. In *Structural Interdependence and Economic Development*, edited by Barna. London: McMillan
Lösch, A. (1940) *Die Raumliche Ordnung der Wirtschaft, Jena*. English translation by W. H. Woglom and W. F. Stolper (1954): *The Economics of Location*. New Haven: Yale University Press
Lowry, I. S. (1964) *A Model of Metropolis*. Santa Monica, CA: Rand Corporation
(1968) Models of urban development. In *Urban Development Models*, edited by G. Hemmens. Special Report 97. Washington DC: Highway Research Board
Luce, R. D. (1959) *Individual Choice Behaviour*. New York: John Wiley
McGill, S. M. (1977) The Lowry model as an input–output model and its extention to incorporate full intersectorial relations. *Regional Studies*, 11, 337–54
McLaughlin, W. A. (1966) *Multi-path System Traffic Assignment Algorithm*. Research Report RB 108. Ontario: Department of Highways
Malisz, B. (1969) Implications of threshold theory for urban and regional planning. *Journal of the Town Planning Institute*, 55, No. 3
March, L. (1972) Quoted from *Urban Systems Study, Demonstrations and Seminars*. Cambridge: Land Use and Built Form Studies
Marx, C. (1847) *Misère de la philosophie*, Paris. Extracts from *Karl Marx, Selected Writings in Sociology and Social Philosophy*, edited by T. B. Bottomore and M. Rubel (1961). C. A. Watts and Co Ltd
Masser, I. (1969) *A Test of Some Models for Predicting Intermetropolitan Movement of Population in England and Wales*. UWP 9. London: Centre for Environmental Studies
Mayberry, J. P. (1970) Structural requirements for abstract-mode models of passenger transportation. In *The Demand for Travel: Theory and Measurement*, edited by R. E. Quant. Lexington, MA: Heath Lexington Books
Mills, E. S. (1969) *Studies in the Structure of the Urban Economy*. Baltimore: Johns Hopkins Press
MINDUR (1986) *Instalacion de un Sistema de Modelos y su Aplicacion a La Victoria, El Concejo y Tejerias*. Caracas: Ministerio de Desarrollo Urbano
Moore, E. F. (1963) The shortest path through a maze. *Proceedings, International Symposium on the Theory of Switching*. Cambridge, MA: Harvard University Press

MTC (1986) *Estudio de la Demanda de Transporte de la Autopista Caracas La Guaira*. Caracas: Ministerio de Transporte y Communicaciones

(1987) *Solicitud de Prestamo al Banco Interamericano de Desarrollo – Subprograma Autopistas*. Caracas: Ministerio de Transporte y Communicaciones

Murchland, J. D. (1969) *The 'Once Through' Method of Finding All Shortest Distances in a Graph from a Single Origin*. LBS-TNT-56.1. PATRAC

Muth, R. F. (1968) Urban residential land and housing markets. In *Issues in urban Economics*, edited by H. S. Perloff and L. Wingo. Baltimore: Johns Hopkins Press

(1969) *Cities and Housing*. Chicago: University of Chicago Press

Owens, S. (1986) *Energy, Planning and Urban Form*. London: Pion

Penz, F. (1983) *Passive Solar Heating in Existing Dwellings: a Study of Potential in the Existing Housing Stock of the United Kingdom*. Cambridge: The Martin Centre for Architectural and Urban Studies

Popper, K. R. (1963) *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge and Kegan Paul

(1966) *The Open Society and its Enemies*. London: Routledge and Kegan Paul

(1972) *Objective Knowledge: An Evolutionary Approach*. London: Oxford University Press

Putman, S. H. (1973) *The Interrelationships of Transport Development and Land Development* (2 volumes). University of Pennsylvania, Philadelphia: Department of City and Regional Planning

(1975a) Urban land use and transportation models: a state of the art summary. *Transportation Research*, 9, 187–202

(1975b) Further results from, and prospects for future research with, the Integrated Transportation and Land Use Model Package (ITLUP). Paper presented at the Annual Conference of the Southern Regional Science Association, Atlanta, Georgia

(1975c) Calibrating a Disaggregated Residential Allocation Model – DRAM: the residential sub-model of the Integrated Transportation and Land Use Model Package (ITLUP). Paper presented at the Regional Science Association, British Section, 8th Annual Conference, London

Reilly, W. J. (1931) *The law of retail gravitation*. New York: Putman and Sons

Rickaby, P. and de la Barra, T. (1987) A theoretical comparison of strategic spatial options for city–regional development, using the TRANUS model. Paper presented at the Conference: Models for Strategic Energy Decisions in an Urban and Regional Context, Stockholm

Rogers, A. (1968) *Matrix Analysis of Inter-Regional Population Growth and Movement*. University of California Press

SELNEC Transportation Study (1971) *The Mathematical Model*. Technical Working Paper No. 5. Town Hall, Manchester

(1972) *Calibration of the Mathematical Model*. Technical Working Paper No. 7. Town Hall, Manchester

Senior, M. L. (1977) Problems in the integration of land use/transportation models. In *Urban Transportation Modelling*, edited by P. Bonsal, Q. Dalvi and P. Hills. London: Abacus Press

Shortreed, J. H. and Wilson, J. (1968) A Minimum Path Algorithm. Unpublished Report, Department of Civil Engineering, University of Waterloo, Ontario

Stewart, J. (1948) Demographic gravitation: evidence and applications. *Sociometry* 11 (Feb.–May)

Stoneman, C. A. (1972) Shortest route through a road network. *Technical Notes, Series A: Computer Topics in Planning*, No. 8. Cambridge: Land Use and Built Form Studies

Supernak, J. (1983) Car availability versus car ownership modelling: theoretical

discussion and empirical findings. *Proceedings of the World Conference on Transport Research*, Hamburg, Germany
UMTA (1977) *UT PS Reference Manual*. US Urban Mass Transit Administration
Van Valkenburg, M. E. (1974) Foreword to *An Introduction to Discrete Systems*, by K. Steiglitz. New York: Wiley
Von Thünen, J. H. (1826) *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*, Hamburg. English translation by C. M. Wartenberg (1966), edited by P. Hall. *Von Thünen's Isolated State*. London: Pergamon Press
Wardrop, J. G. (1952) Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers*, Part II, 1, 325–78
Weber, A. (1909) *Uber den Standort der Industrien*, Tübingen. English translation with introduction and notes by C. J. Friedrich (1929), *Alfred Weber's Theory of the Location of Industries*. Chicago: University of Chicago Press
Williams, H. C. W. L. (1977a) On the formulation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning A*, 9, 285–344
(1977b) Generation of consistent travel demand models and user-benefit measures. In *Urban Transportation Planning*, edited by P. Bonsall, W. Dalvi and P. Hills. London: Abacus Press
Wilson, A. G. (1967) A statistical theory of spatial distribution models. *Transportation Research*, 1, 253
(1970) *Entropy in Urban and Regional Modelling*. London: Pion
(1974) *Urban and Regional Models in Geography and Planning*. Chichester: John Wiley
Wilson, A. G., Hawkins, A. G., Hills, G. J. and Wagon, D. J. (1969) Calibrating and testing the SELNEC transport model. *Regional Studies*, 3, 337–50
Wingo, L. (1961) *Transportation and Urban Land*. Baltimore: Johns Hopkins Press
Wootton, H. J. and Pick, G. W. (1967) A model for trips generated by households. *Journal of Transport Economics and Policy*, 1, 137–53
Zipf, G. K. (1949) *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley

Index