A UNIFYING FRAMEWORK FOR PUBLIC FACILITY LOCATION PROBLEMS

Giorgio Leonardi
International Institute for Applied Systems Analysis, Austria

RR-81-28
November 1981

Reprinted from *Environment and Planning A*, volume 13 (1981)

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
Laxenburg, Austria

Research Reports, which record research conducted at IIASA, are independently reviewed before publication. However, the views and opinions they express are not necessarily those of the Institute or the National Member Organizations that support it.

Reprinted with permission from *Environment and Planning A* 13: 1001–1028, 1085–1108, 1981.

FOREWORD

The public provision of urban facilities and services often takes the form of a few central supply points serving a large number of spatially dispersed demand points: for example, hospitals, schools, libraries, and emergency services such as fire and police. A fundamental characteristic of such systems is the spatial separation between suppliers and consumers. No market signals exist to identify efficient and inefficient geographical arrangements; thus, the location problem is one that arises in both East and West, in planned and in market economies.

This problem is being studied at IIASA by the Public Facility Location Task, which started in 1979. The expected results of this Task are a comprehensive state-of-the-art survey of current theories and applications, an established network of international contacts among scholars and institutions in different countries, a framework for comparison, unification, and generalization of existing approaches, as well as the formulation of new problems and approaches in the field of optimal location theory.

This report consists of two parts: the first is a nontechnical description of the proposed general framework for analyzing location problems, the second describes mathematical models for static single-service facility-location problems and their possible extensions and improvements.

ANDREI ROGERS
Chairman
Human Settlements and Services Area

Environment and Planning A, 1981, volume 13, pages 1001-1028

A unifying framework for public facility location problems—part 1: A critical overview and some unsolved problems

G Leonardi
International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria
Received 19 June 1980

Abstract. This paper, a condensed report of the present state of the work in the Public Facility Location Task (formerly the Normative Location Modeling Task) at IIASA, has three main aims: first, to build a general framework for location problems; second, to use this framework to unify existing location models; and, third, to use the framework to develop new, more general, and more meaningful location models. Suggestions are also given on how to introduce multiple services and multiple time periods in location problems. The multiactivity dynamic location models that this perspective generates is the subject of future research in the Public Facility Location Task.

This first part of the paper gives a nontechnical description of the proposed general framework for analyzing location problems. The second part will describe mathematical models for static, single-service, facility location problems and their possible extensions and improvements, and will appear in the next issue.

1 Introduction

The Public Facility Location (PFL) Task, formerly called the Normative Location Modeling Task, was undertaken in a rather exploratory way in October 1979, within the Human Settlements and Services Area at IIASA. In spite of the esoteric-sounding terminology, the PFL problem is one that is experienced daily and can be rephrased as: How can the location of public services and facilities be planned in an optimal way? A first step is to put aside the problem of the precise definition of such terms as 'planning', 'location', 'public', 'service', and 'optimal', and, instead, focus attention on operational tools, a topic already the subject of a vast amount of literature.

To begin a study of location modeling the following general goals are useful:
(1) a comprehensive review of the existing theoretical and applied literature on optimal location should be made;
(2) after this has been done a state-of-the-art review of all work on optimal location modeling should be assembled; and
(3) simultaneously an international communication network among scientists working on location problems should be built.
Although the above goals have actually been implemented and have been found useful, as the work progresses, more precise definitions are soon required:
(a) By far the greatest unsolved problem in this field of study is the lack of interdisciplinary work. Although this statement sounds just as general as the three main goals, it is not. A glance at the literature on location problems is enough to realize that most approaches and results are as diverse and scattered as the different trainings and backgrounds of their authors. It is for this reason that international cooperation among scholars studying location problems is important.
(b) In order to establish an efficient interdisciplinary study, general unifying frameworks are required for location problems. This seems to be the most promising way to approach the state-of-the-art review on the subject, and the most likely way for IIASA to make an original theoretical contribution.
(c) If optimal location problems are not to be doomed to the realm of mathematical skill games, they must be related to the more general and realistic problem of planning

optimal locational structures for interacting urban activity systems. This is likely to be the most promising applied IIASA contribution.

With these goals in mind, work has begun at IIASA, and some results have been achieved. The purpose of this paper is to present this work, first by stating the limitations and shortcomings of the present state of the research and then by pointing to the main achievements and suggested approaches for further research.

There are four main shortcomings to this research. First, a general theory has yet to be developed, although some unifying issues have been exploited. However, it must be said that the path towards unification is not really difficult; it simply requires some time and effort, and it proves itself to be a fruitful one.

Second, some important approaches to the study of location modeling are lacking, as a result of limitations in time and knowledge. Among them, mention should be made of the multiobjective approach, the voting decision procedures, and the use of pricing policies (parallel to, or alternative to, physical-stocks planning).

Third, a deeper analysis of the relevant welfare theory behind the optimization models has not been carried out. This seems indeed to be one of the most promising issues for future research, both on a microeconomic and on a macroeconomic scale.

Fourth, some algorithmic problems still require a rigorous solution, although heuristic tools that give good solutions are available. This is another strand of future research, and a challenge for applied mathematicians.

The main achievements and suggested approaches for future research can be categorized into three groups beginning with the physical interpretation of results, solutions, and main properties. This approach has always been kept in mind, and easily understandable rules-of-thumb have been sought and suggested when possible. The general direction implied here is that a qualitative understanding of the structure of the problem is sometimes more useful than being able to generate numerical solutions. Although this issue is listed among the achievements, it must be said that only the tip of the iceberg has been scratched, and a considerable amount of work is needed along this path.

The second main asset of this study is that all the newly proposed models have been built on existing ones that have been included as special cases. This is a suggested standard for future research; nothing has to be thrown away, and everything can be used as a start for a generalization. This is not just a theoretical issue, as it may appear to be at first sight. Including existing problems as special cases usually implies being able to generalize special problems, as well as solution techniques.

The last main contribution of this work is the more realistic assumptions that have been introduced for users' behavior. This approach is also just in the beginning stage, and the problem behind it is worth further theoretical and empirical research.

2 General issues

2.1 *The facility location system*

Broadly speaking, the general aim of a location pattern, for a given set of public facilities, is twofold:

1. to be as near as possible to the demand, in order to reduce transport costs; and
2. to keep the cost of establishing the facilities as low as possible, both by choosing low-cost locations and by reducing the number of facilities to be established.

Since these two goals are usually in conflict, some trade-off has to be found between them. The need for such a trade-off is the reason why nontrivial location problems exist.

In order to build a general framework, it is useful to split the location problem into two subproblems.

(a) The first is the problem of *allocation*, where the transportation pattern between the demand locations and the service facilities is decided. The allocation problem is thus mainly concerned with the first goal of reducing transport costs.
(b) The second is the problem of *facility location*, where the locations of the facilities are chosen. The location problem is thus mainly concerned with the second goal of low-cost location.

Of course, the allocation and location subproblems are related and cannot be solved separately. However, this does not mean that they are both controlled by the same decisionmaker, nor that the two possible decisionmakers agree. Figure 1 shows the way in which the two subproblems interact in the most general case.

Two formal operators have been introduced: the *locator* and the *allocator*. They are defined below:

Locator's decision—the locator operates as follows:
(a) compares the current location and size of existing facilities with the demand allocated to them;
(b) evaluates the costs and benefits associated with the current state of the system and with its possible changes; and
(c) generates a new pattern of sizes and location of facilities, both by establishing new facilities and by demolishing old ones. (Of course, building and demolition costs are taken into account at this stage.)

Allocator's decision—the allocator operates as follows:
(a) compares the potential demand from each demand location, the existing size and location of facilities, and the transport costs (not necessarily measured in terms of money) between the demand and the facility location; and
(b) generates a pattern of trips between the demand and facility locations, taking into account both transport costs and available capacity.

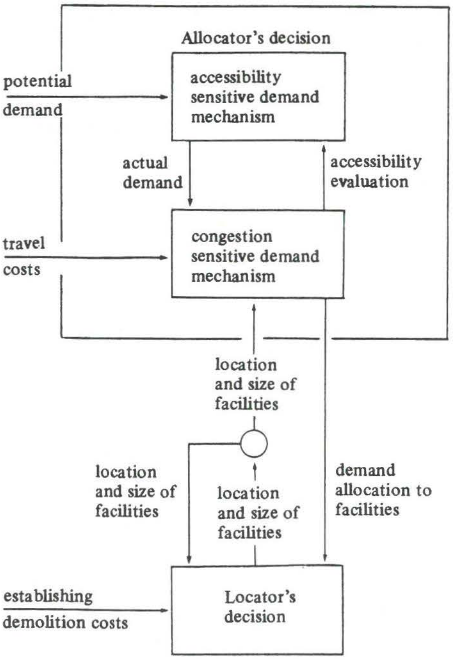
Allocator's decision
potential demand
accessibility sensitive demand mechanism
actual demand
accessibility evaluation
travel costs
congestion sensitive demand mechanism
location and size of facilities
location and size of facilities
location and size of facilities
demand allocation to facilities
establishing demolition costs
Locator's decision

Figure 1. The general location allocation system.

Although it has already been stated, it is worth stressing again that location and allocation are not necessarily (nor usually) decided by the same decisionmaker. More specifically, it often happens that although location may have a single decisionmaker (for example, a planning authority or a firm) allocation may be decided by customers. In the latter case, the resulting trip pattern is the pooled output of many individual decisions. Such an aggregate of microscopic decision units will still be called an 'allocator' operator, although it is in no way necessarily implied that it has a rationally consistent overall behavior.

In order to have a better understanding of how the allocator works, it is useful to make a further division into two subsystems. The first of these is the *accessibility-sensitive demand mechanism.* In many services the total actual demand cannot be estimated beforehand; rather, it depends on the location, size, and proximity of the facilities. Usually, the demand will increase with the accessibility of the service and is, therefore, accessibility-sensitive. The accessibility-sensitive demand mechanism receives as inputs the potential demand (that is, the maximum demand which can be generated) and the evaluation of the accessibility to the service for each demand location. It then generates the actual demand from each demand location, as a fraction of the potential demand, according to some accessibility nondecreasing function.

The second subsystem necessary in the allocator's decision process is the *congestion-sensitive demand mechanism.* Just as demand can be generated by high accessibility to the service, it can also be discouraged by overcrowded facilities. An example where both mechanisms operate is leisure and recreational facilities. Usually, the higher the availability and nearness of swimming pools, the higher will be the number of people using them. On the other hand, if a swimming pool is overcrowded, people will tend to look for another one, or maybe go back home. Therefore, the congestion-sensitive demand mechanism receives as inputs the actual demand from each location, the location and size of facilities in each location, and the transport costs between demand and facility locations. It then generates the evaluation of accessibility from each demand location (accounting both for nearness and for congestion) and the travel pattern between demand and facility locations.

The interaction between the two allocator subsystems is perhaps better understood if rephrased in dynamic terms. Starting with some a priori estimate of accessibility, the accessibility-sensitive demand mechanism generates a demand. This demand is fed into the congestion-sensitive demand mechanism, which then allocates it to facilities, evaluates the resulting congestion in each facility, accounts for this in evaluating a new estimate of accessibility, and sends this information as a feedback to the accessibility-sensitive demand mechanism. A new actual demand, depending on the new congestion-reduced accessibility evaluation, is thus generated, and the same process is repeated over and over again, until some equilibrium state is reached, provided it exists.

2.2 *The service mechanism*

2.2.1 *Introduction.* The general qualitative framework defined in section 2.1 is wide enough to include most service location problems. However, it does not give answers to such questions as: Who are the locators and the allocators? How much do they agree or disagree? Who (or what) is traveling, where is he (or it) traveling to and why? Who is paying for travel costs? Who is paying for establishing facilities?

Most of these questions can be given a definite answer only by introducing a definite kind of service. However, some useless generality can be eliminated by assuming that all public services and facilities share the following two reasonable features and properties:

(a) The locator is a public authority; he basically agrees with the users and their goals (the case of the 'bad tyrant' will be neglected), and sometimes he may also be assumed to be the user himself.

(b) The locator, a public authority, pays for establishing the service facilities.

Note that no special assumption for the allocator has been introduced. This is because of the varying behavior of the allocator which is the very reason different service mechanisms can be found. The distinction between the public decisionmaker and the users may be sharp or weak, as shown by the following extreme cases. When the same decisionmaker controls both the locator and the allocator, that is, when there is no difference between public authorities and users, *all costs* (travel and establishing costs) are paid by the users. An example of how this may happen is the US primary school system, where users pay location-dependent taxes and at the same time control to some extent the location, size, and allocation of their own school facilities. The US case does not hold in most European countries, where the very opposite can sometimes be found. In Italy, for instance, overhead (location-independent) taxes are paid for primary schools, but location and allocation are centrally planned by a public authority, so that the resulting allocations may sometimes be in conflict with users' preferences.

The above examples focus attention on two main families of service mechanisms: services where the same decisionmaker controls location and allocation, referred to in this paper as *delivery systems*, and services where location and allocation are controlled by different decisionmakers, referred to as *users-attracting systems*.

2.2.2 *Delivery systems.* A delivery system is defined in the following way.

(1) The same decisionmaker (usually a public authority and/or an agency) controls location of facilities and allocation of services to users.

(2) Users *do not travel*; that is, service is delivered from the facilities to the users, and transport costs *are not* charged to the users.

(3) *Every cost* (including transportation) is paid by the same public decisionmaker.

There are many examples of delivery systems of which three are mentioned here. Fire emergency services fulfill the three requirements, as long as no bill has to be paid for fire services. Police patrol systems is a second example of delivery systems. Health care emergency (or ambulance) systems, can also be considered as delivery systems, as long as the decision to use them is not affected by transport costs and is therefore price-insensitive.

The main distinguishing feature of delivery systems is the direction and kind of transport flows. Transport always takes place from facility locations to demand locations, and users *do not travel*; they are served at home. A delivery system is the simplest location problem one can conceive, since no conflicting goals arise (except for the usual transport-location trade-off problem). No user's behavior model is needed, since users do not behave at all in such systems (the reasonable assumption of accessibility and congestion insensitive demand is implied here for such systems—hopefully, people do not cause fires just because they live near to a fire station), and every cost is paid (or can be assumed to be paid) by the same decisionmaker. Figure 2 shows the typical direction of transport flows in a delivery system.

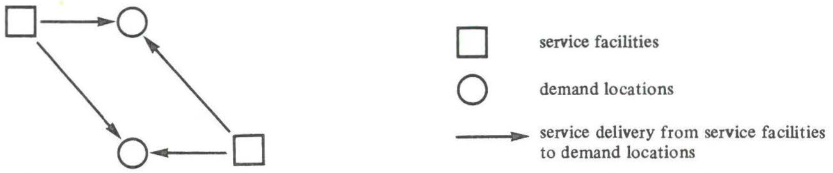
□ service facilities
○ demand locations
⟶ service delivery from service facilities to demand locations

Figure 2. A typical delivery-system flow pattern.

It may be useful to point out that a delivery system closely resembles the classical 'plant-location' or 'warehouse-location' problems used in private-sector locational decisions and found in the operations research and management science literature (Balinski, 1961; Efroymson and Ray, 1966; ReVelle and Rojeski, 1970; Hansen and Kaufman, 1976; Erlenkotter, 1978; van Roy and Erlenkotter, 1980). This means that most (perhaps all) delivery-system location problems can be solved by well known and reasonably simple existing techniques.

2.2.3 *Users-attracting systems.* A users-attracting system is the very opposite of a delivery system; it is the appropriate model for most meaningful urban service systems, and it also raises unsolved problems and needs further research. The typical service mechanism of a users-attracting system is defined in the following way:
(1) Location and allocation are controlled by different decisionmakers. For our purposes, it can be assumed that location is controlled by a planner (usually a public authority or agency) whereas allocation is controlled by users. (Users are not single decisionmakers, they are an aggregate of possibly goal-conflicting people. Let it be provisionally assumed, however, that conflicting problems among users can be neglected, and 'average' users' goals can be defined.)
(2) Users travel; that is, they have to go to the facility locations in order to be served. This also means they have to pay for traveling.
(3) Although transport costs are charged to the users (recall that transport costs are not necessarily paid in terms of money), costs for establishing facilities are paid by the planner or public decisionmaker.

Three good examples of users-attracting systems are some schools, some health care systems, and cultural and recreational services. High schools, for instance, often meet all three of the above definitions, although the first definition needs some care when it is applied to primary schools, since it is not always clear who are the users and what goals they are pursuing. In the author's opinion, children are the users; but it is well known that children do not like to go to school, although public educational authorities force them to go. This is a typical conflicting-goals problem. In some other people's opinion parents are the users, and in this case a user–planner general agreement is more likely to be found.

With the exceptions only of emergency ambulance systems and of home health care delivery systems, the health care system in many countries meets the three definitions given above. Indeed, the health care system poses new modeling problems, since it exhibits a multilevel structure as far as the patient—or user—is concerned. A typical patient history could be the progression from a day-hospital for a check-up to a specialized hospital where the patient was found to have cancer, then to a specialized surgical hospital where he was operated on, after which he was sent to a rehabilitation center until there was a full recovery. If he is lucky enough, he will find all the facilities he needs in the same locations. But this is unlikely, since, although general purpose facilities (for example, day-hospitals) may be fairly scattered, specialized or infrequent treatments tend to be concentrated in a few locations (for example, rehabilitation centers). The health care facilities optimal location must take this multilevel structure and behavior into account, in order to evaluate *not only* accessibility from demand locations to facility locations, *but also* accessibility within facilities belonging to different levels.

Public libraries, theaters, and swimming pools are several examples of cultural and recreational services where the travel pattern resulting from users' behavior is usually far from a planned one. Different tastes may lead to choices which disagree with distance minimizing. Unlike the delivery systems, in the users-attracting systems every possible complication can arise (and usually does). Location goals (or planner goals) and allocation goals (or user goals) may be different and conflicting, although the public facility assumption usually implies they are not. A model for users' behavior is needed, since usually such simple rules as nearest-facility allocation are unrealistic and do not fit the actual behavior. Accessibility and congestion sensitiveness may also be required, as in the case of recreational services. The typical direction of transport flows in a users-attracting system is shown in figure 3.

It may be useful to point out that models for users-attracting systems are hardly found in the operations research (OR) and management science literature, except for special cases, which can be (or are forced to be) reduced to the usual plant-location form. On the other hand, the problem of modeling location systems based on actual users' behavior appears in most regional science and urban geography literature (for example, Lowry, 1964; Harris, 1964; Huff, 1964; 1966; Lakshmanan and Hansen, 1965; Wilson, 1970; 1974; 1976; Coelho and Wilson, 1976; 1977; Leonardi, 1978; 1979a; Harris and Wilson, 1978). Most of these models are based on the so-called gravity, or spatial interactions assumption, according to which the number of trips between demand and facility locations is proportional to a smoothly decreasing function of distance or travel cost (other imposed constraints being met).

Although the above models are based on more general and sensible physical assumptions, as compared to the usual OR 'plant-location' models, the development of mathematical tools and algorithms have not been as good as in the OR field. This is an area where some unsolved problems can still be found, and further research is needed.

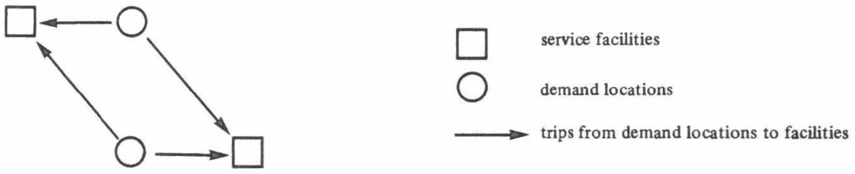

Figure 3. A typical users-attracting system flow pattern.

2.3 *The users' behavior*

2.3.1 *Introduction.* The purpose of section 2.3 is to introduce some specific assumptions and models for users' trip-making and facility-choosing behavior. Therefore, from now on reference will be made only to the users-attracting systems. Here again it can be said that, although each kind of service is associated with a different user's behavior, many common features in the models and techniques can be found. The main issues leading towards some unifying results can be summarized into three groups.

The problem of relating users' choices to some optimizing behavior is the first main issue. Although a completely random-choice behavior is possible, in most meaningful cases some regularities can be found, which lead to the assumption that users consistently choose locations with some overall optimization criterion. Gravity and spatial interaction models play an important role in this field, since they have both good empirical testing and optimization interpretations.

The second issue is the problem of defining a suitable general form for an accessibility-sensitive demand mechanism. This is a problem seldom found both in OR and in regional science literature. It is therefore a good topic for further research, and a most promising one, since many public services exhibit accessibility-sensitive demand.

The third problem is that of defining a suitable, general form for a congestion-sensitive demand mechanism. Of course, congestion mechanisms are the main subject of queuing theory. But the queuing theory approach is micro and is therefore unlikely to be useful to solve problems of realistic size (that is, with many service locations). Some simpler models are therefore needed, perhaps based on more aggregate assumptions and variables than queuing models.

In the following subsections a general, nontechnical description of ways to solve the above problems is given, and some possible further generalizations are suggested.

2.3.2 *Planned versus unplanned allocation.* Although it sounds like a contradiction (and in the opinion of the author it is, to some extent), it is worth considering the case of users-attracting systems where allocation is centrally planned, although the users pay for travel costs. This is not necessarily a strong inconsistency with the public nature of the services to be located. If, for instance, users have a general tendency to minimize travel costs, but may also wish to trade off other costs or benefits (such as congestion or quality of service), although a public decisionmaker strictly minimizes travel cost, it cannot be said that he is really being inconsistent with users' goals. He is just being narrow-minded and uninformed on some details of the users' preferences, although he is aware of some of their main and simplest tendencies (such as travel cost minimizing). Since he is unable to predict what users will do when facilities will be provided, he just *imposes* on them a behavior according to *his simplified version* of their own goals. This approach is somewhat justified (although not necessarily to be agreed with) when there is no difference between users and public decisionmakers, in either goals or costs to be paid, or when the public decisionmaker has enough normative power to effectively impose the allocation rule. The US–Italian primary school location–allocation examples are relevant again here.

As already stated, there is no difference between such problems and classical plant-location problems, as far as mathematical models and solution algorithms are concerned. The induced users' behavior follows the simple nearest-facility allocation rule depicted in figure 4.

Let it now be assumed that at least one of the two justifications for the approach of the decisionmaker no longer holds true, or is likely to be unfair. It may happen, for instance, that trade-offs between travel costs and other users' costs and benefits are too relevant to be neglected. Obvious examples are health care (nonemergency) services, most kinds of schools, and cultural and recreational facilities. It may also happen that the public planner has no real normative control on the way users allocate themselves to facilities. He is just hoping they will go to the nearest facilities, and he makes his plan accordingly. In this case, allocation should not be referred to as 'planned'. What the public decisionmaker is actually doing is using a bad, unreliable users' behavior model.

Many more examples could be given, which regretably raise complicated social and political issues, outside the scope of this paper. Let it be assumed, therefore, that there are enough intuitive reasons to focus our interest on location systems where allocation is left to the users' unplanned behavior. These systems will be referred to as 'unplanned allocation systems'.

Without much loss of generality, let it also be assumed that, other things being equal, users tend to minimize travel cost. It then follows that possible deviations from this tendency may result if some of the 'other things' are not perfectly 'equal'.

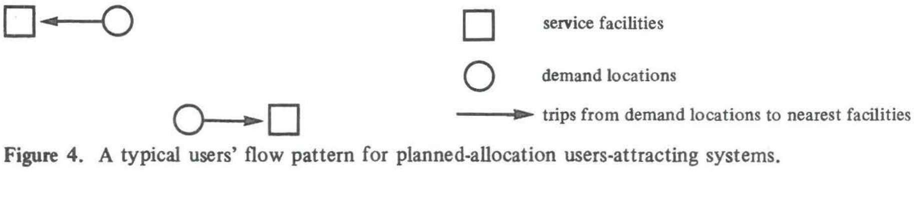

Figure 4. A typical users' flow pattern for planned-allocation users-attracting systems.

If the overall result of such a process is examined, one can expect to find a frequency of users' trips that decreases with the cost of traveling from demand location to facilities. The usual trip pattern resulting from this behavior looks like the one shown in figure 5. That is, although most trips are to the nearest facility, a sizeable amount are to a further facility. The pattern shown in figure 5 is an intermediate one between two possible extremes:

(a) All flows go to the nearest facilities, which then reduces to the already discussed planned-allocation problem.

(b) Flows are evenly split among facilities, which is equivalent to a users' behavior totally insensitive to travel cost.

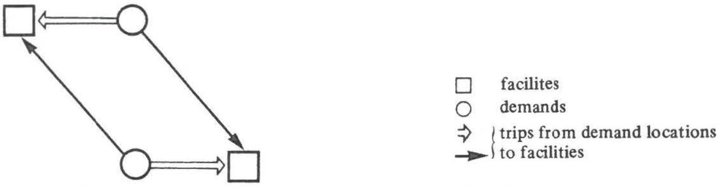

Figure 5. A typical users' flow pattern for unplanned-allocation users-attracting systems.

2.3.3 *Spatial-interaction-based allocation models.* This section focuses on possible useful ways to handle the unplanned-allocation case discussed in section 2.3.1. Provided users are assumed to have different tastes, value judgments, and cost–benefit evaluations for each facility, their behavior can be modeled by using one of the following two approaches.

1. *The disaggregate approach*, where a complete list of all users in each location is kept, and the utility functions (or at least the preference ordering relations) on alternative facilities are defined for each user in the list. This approach has been recently proposed by Hanjoul (1980).

2. *The aggregate approach*, which is useful when the list of users becomes very large and it is impractical to keep track of every user and hopeless to define so many preference orders. Some general regularities in aggregate behavior are usually found, when the overall result of superimposing all these different preference orders is observed. The frequency of choice usually decreases with travel costs. If needed, or if available data make it possible, other relevant explanatory variables can be introduced. Such observed data can be plotted and fitted to some curve (when a good fit is usually found).

It is obvious that the first approach can be used only when the number of different users, or preference orders, in the list is small. It is, however, interesting when microbehavior exploration is the main concern. But for operational purposes the aggregate approach seems better, when there are so many users that they cannot be listed one by one. (This is, by the way, the usual case with real public facilities.)

Approach 2 gives rise to what are usually referred to as 'gravity', or 'spatial-interaction' models. In loose terms (more technical definitions will be given in later sections), the main feature of a spatial-interaction model is to replace the very sharp assumption of travel-cost minimizing with the smoother assumption of space-discount behavior. Whereas the travel-cost-minimizing user places infinite value on the nearest facility, the space-discounting user ranks all facilities in a distance-decreasing order, the distance-decrease shape being given by a curve similar to those shown in figure 6. Moreover, space-discounting users are assumed to be stochastic, and choose facilities with probabilities proportional to the space-discount factor (other possible constraints being met). Stochastic behavior here is merely a model of one's ignorance, since detailed information on each user has been lost after the aggregation process[1].

A useful property of the family of curves shown in figure 6 is that it includes the nearest-facility allocation rule as a special case (curve 1). It also includes the opposite limiting case, that is, the random (distance or travel cost independent) allocation rule (horizontal line 4). In real problems, however, cases will be found between these two extremes: 2 and 3.

Although gravity or spatial-interaction models were first developed empirically, many alternative theoretical justifications for them have been proposed in the last ten years, so that what once seemed to be slightly more than a rule-of-thumb model has become a topic worthy of consideration by theoretical economists and geographers, mathematicians, and statisticians, besides regional scientists. Such theoretical works range from the classical entropy maximizing approach (Wilson, 1970), which is an aggregate one, to disaggregate stochastic-choice models, among which the logit model is the best known (McFadden, 1974a; 1974b), and to models that derive macroeconomic interpretations by aggregating random-choice models, such as in the consumer surplus-maximizing approach (Neuburger, 1971; Williams, 1977; Coelho and Williams, 1978; Coelho and Wilson, 1976), or in the accessibility-maximizing approach (Leonardi, 1973; 1975; 1978; Williams and Senior, 1978).

In spite of the seemingly strong differences among these approaches, it is surprising how they all give rise to the same models, as far as their mathematical form is concerned. This is not really surprising, if one considers the fact that the logit model, which is supposed to have definite roots in random utility theory (being a typical 'micro' model), is merely a multinomial logistic distribution. On the other hand, the entropy maximizing approach, which has been developed in statistical mechanics (and makes no assumptions on the microscopic behavior of the system under study, but just poses some weak constraints at the aggregate level) gives rise to the same multinomial logistic distribution. (The Boltzmann probability distribution used in thermodynamics and the Fermi-Dirac distribution of quantum mechanics are nothing but special forms of logit models.)

Besides the theoretical significance of this general consistency between different interpretations, it is operationally comfortable to know that one has to work with just one mathematical formulation. Although the subject will be treated in more detail in later sections, it is worthwhile at this point to give the general form for a spatial

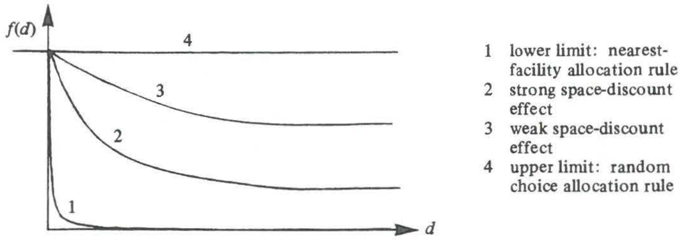

Figure 6. General shape for a family of space-discount functions.

[1] A microeconomic stochastic behavior might also be introduced, if each user is assumed to have a probability distribution on utilities assigned to each facility. This is basically what random utility theory is. But luckily enough random-utility based models are indistinguishable from gravity models at the macrolevel. They are actually a possible alternative interpretation of the same models, rather than new ones.

interaction model, which is as follows:

$$S_{ij} = G_i \frac{f(C_{ij})W_j}{\sum_j f(C_{ij})W_j} \, , \tag{1}$$

where

i, j are subscripts labeling the locations of demand and facilities, respectively,
G_i is the total demand for service generated in i per unit time,
W_j is a measure of attractiveness of facility in j,
C_{ij} is the total cost associated with a displacement from i to j, measured in appropriate units,
$f(\cdot)$ is a space discount function, of the kind discussed above, and
S_{ij} is the number of customers living in location i and using the facility location j.

The space-discount function is often assumed to have some special form, like a negative exponential, or a negative power function, but none of these assumptions are needed to analyze the general structure of a spatial-interaction model. On the contrary, terms such as 'generation' and 'attractiveness' need a better definition, since they are related to the problems of modeling sensitiveness of demand both to travel cost or accessibility and to congestion or overcrowding of facilities.

2.3.4 *Accessibility-sensitive demand.* Most existing public facility location models assume both allocation of customers to the nearest facility and insensitiveness of demand to accessibility to the service. But, as already discussed in the introductory sections, this assumption is not appropriate for many services, where the total demand cannot be estimated independently of the size and location of facilities. In other words, demand is induced by the provision of the service, and usually the easier facilities can be reached, the higher will be the demand. In loose economic terms, it may be said that travel cost acts as a price to be paid in order to use the service, and total demand is nonincreasing with this cost. Although the sensitive-demand problem has been stated in market-like terms, it is in no way related to market economies rather than to planned economies. Travel costs must be paid everywhere, since they depend on the existence of space distance, and are not necessarily measured in terms of money. (In most urban settlements travel time is usually the best measure of travel cost.)

In order to define a general structure for an accessibility-sensitive-demand model, some definite measures of accessibility have to be introduced. The three main possibilities are listed below.

(a) Accessibility is measured by means of either the nearest facility travel cost or the arithmetic mean of travel costs to all facilities. Both measures tacitly assume that customers agree on evaluating distances according to a minimum distance, or an arithmetic mean basis. These assumptions are usually inconsistent with a general spatial-interaction behavior.

(b) Accessibility is a measure of users' benefit consistent with a spatial-interaction behavior. Since according to most spatial interaction data, users seem to apply a definite distance-decreasing discount factor on facilities, the most natural measure of accessibility seems to be a sum of the capacity (or attractiveness) of all service facilities, each one discounted with its own space-discount factor. Such a measure is closely related to early concepts of social physics and regional analysis (like potentials and related concepts). [See Steward (1948) and Vickerman (1974) for a review.] In its modern form it has been introduced by Hansen (1959), and developed by Ingram (1971), Weibull (1976), Leonardi (1976), Smith (1976), Davidson (1977), and Sheppard (1979). Its general form is:

$$A = \sum_j f(C_j)W_j \, , \tag{2}$$

where

A is the measure of accessibility to the service from a given demand location,

W_j is a weight measuring attractiveness of a facility in j,

C_j is the cost of traveling from the demand location to the facility in j, and

$f(\cdot)$ is a space-discount function.

(c) Accessibility is measured by an 'average' travel cost, where the averaging operator is consistent with a spatial-interaction behavior. Since users perceive distance by means of a space discount function, which is averaged to build an accessibility measure [see (b) above], it is natural to obtain a measure of 'average' travel cost by applying the inverse of the space-discount function to accessibility. That is,

$$\overline{C} = f^{-1}(A) , \tag{3}$$

where

$\overline{C}$ is the average travel cost from a given demand location having accessibility A defined by equation (2), and

$f^{-1}(\cdot)$ is the inverse function of $f(\cdot)$.

Average travel cost as computed by equation (3) is the only possible average ensuring consistency in problems of aggregating and disaggregating spatial-choice models. It has been discussed in depth by many authors (usually for special functional forms), among them Wilson (1974), Williams (1977), and Leonardi (1979b).

Two general forms are possible for an accessibility-sensitive-demand generation model depending on whether accessibility is measured by some actual 'accessibility' index, such as given by equation (2), or by some measure of 'average' (including minimum distance) cost. The first general form is an *accessibility-increasing demand curve*, such as the one shown in figure 7. In general, demand G will increase up to a maximum value $\overline{P}$, corresponding to the maximum value of A, which is reached when all travel costs are zero (that is, space disappears). Furthermore, generated demand G will be bounded from above by a maximum potential demand P, which is always finite (even though possibly very large) if the total population living in the demand location is finite.

The second general form for an accessibility-sensitive-demand model is an *average-cost-decreasing demand curve*, such as the one shown in figure 8. In general, demand G will reach its (physically feasible) maximum $\overline{P}$ for zero cost, and then decrease as travel cost increases. The upper bound P has the same meaning as before.

Of course there is no real difference between the two formulations given above, since the two demand curves can be mapped one on the other. The choice of the best formulation is a matter of convenience.

Perhaps it is also worth mentioning that, just as nearest-facility travel cost is included in 'average costs' many special measures are incorporated in 'accessibility

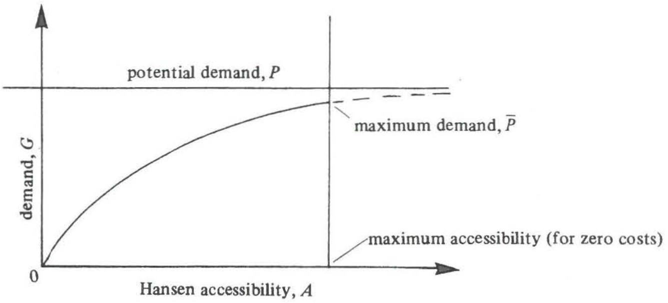

Figure 7. An accessibility-increasing demand curve.

measures' as defined by equation (2), including the widely used 'coverage-based' accessibility measure. This measure is defined as the number (or the total capacity, or any measure proportional to them) of facilities which can be reached within a given maximum travel-cost range. This is a special case of equation (2), where $f(\cdot)$ is a step function, such as the one shown in figure 9.

However, a step function is not the best one to be used for elastic-demand models, since the way it evaluates distance is typical of emergency services, which usually (and hopefully) have inelastic demand.

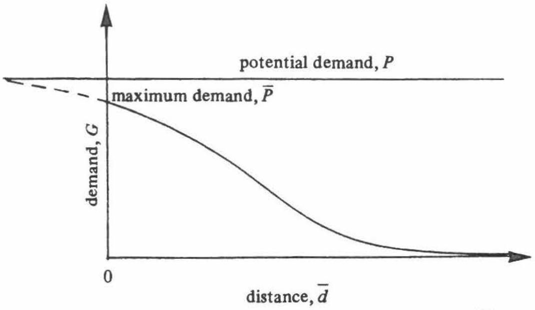

Figure 8. An average-cost-decreasing demand curve ($\bar{d}$ is either average distance or else nearest facility distance).

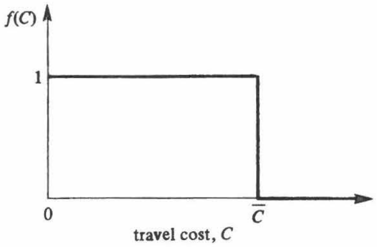

Figure 9. Step space discount function for 'coverage' accessibility measures ($\bar{C}$ is the maximum range).

2.3.5 *Congestion-sensitive demand.* As with accessibility-sensitive demand, most facility location models ignore possible demand sensitiveness to congestion. Indeed, the best-known models in the literature on static optimal location modeling are 'uncapacitated', which means they assume a total capacity that always matches total demand.

Just like travel cost, congestion is a cost that all customers have to pay (usually in nonmonetary units) in order to be served, no matter what economic system rules their country. Although congestion costs are not as unavoidable as travel costs since there is no real physical bound on building new facilities, capacity expansion usually does not solve the problem. This is true mainly when the accessibility-sensitive-demand mechanism is also at work, since increase in capacity increases accessibility, which in turn generates new demand. The joint effect of both mechanisms may therefore result in increased congestion because of increased capacity. This frequently happens with many services, not necessarily of a recreational nature, such as nonemergency health care facilities. It always happens when capacity expansion is decided locally, without an overall look at the whole system of locations.

The simplest congestion-sensitive-demand model (although by no means the only possible one) is a linear feedback signal that is proportional to the difference between total capacity and attracted demand, which changes the value of the attractiveness weights for each facility. (It is assumed that attracted demand may never be greater than total capacity, by definition.) The resulting weights give a new value for accessibilities, which in turn generate different values for total demand, and so on, over and over again, until some equilibrium (if any) is reached. This two-step mechanism can be given the following, simple mathematical formulation.

1. Congestion-sensitive attractiveness weights are evaluated according to the formula

$$W_j = \alpha_j(X_j - D_j)\,, \tag{4}$$

where

X_j is the capacity of facility in location j,
D_j is the total demand attracted in location j, and
α_j are given constants, typical of each location.

2. The new accessibility resulting from the new weights is evaluated by means of equation (2).

$$A(X-D) = \sum_j f_j(C_j)W_j \tag{5}$$

where

$A(\cdot)$ is the accessibility measure, expressed as a function of the differences between capacity and demand.

Although accessibility has been used in the above formulation, of course a similar formulation can be given in terms of average travel cost.

It may be worth noting that congestion found at the destination is not the only possible kind of congestion evaluated by customers. Congestion found on the road links used to reach the destination also determines the choice. However, a model embedding both optimal location and allocation and optimal traffic assignment goes beyond the scope of this paper, and will be the subject for future research. Such models could be developed building upon the work of Evans (1976) and Boyce and LeBlanc (1979). Here for simplicity it will be assumed that the traffic generated by the set of facilities under consideration is small enough to leave the prevailing traffic conditions unaltered. This assumption is reasonable when a single type of facility is considered, but it may become unrealistic for a multiactivity location problem.

2.3.6 *Overall structure of the accessibility-congestion-sensitive allocation system.* Now that all of its subsystems have been defined, the overall allocation system can be assembled. Its general structure is shown in figure 10.

The contents and meaning of each subsystem need no further explanation, since they have been described in detail in the preceding sections. The flows linking the subsystems are worth some explanation, however. Taken as a whole, the system receives two inputs, the potential demand $\boldsymbol{P}$ and the capacity $\boldsymbol{X}$, and gives the allocation matrix $\mathbf{S}$ as an output. Whereas the input $\boldsymbol{P}$ is totally exogenous, possibly given by some population forecast model, the input $\boldsymbol{X}$ is a control vector, whose choice determines size and location of facilities. The output $\boldsymbol{S}$ will be subject to evaluation in the decision process, together with the control vector $\boldsymbol{X}$, since benefits and costs will depend on them. Two main feed-back loops can be seen in the inner flows.

(a) The main loop links demand generation, allocation, and congestion evaluation. The demand generation subsystem (or accessibility-sensitive-demand model) receives an accessibility estimate $\boldsymbol{A}$, generates a demand $\boldsymbol{G}$ and sends it as an input to the allocation subsystem. This subsystem (using also the current estimate of attractiveness $\boldsymbol{W}$) allocates the generated demand $\boldsymbol{G}$ to facilities, by means of a general spatial-interaction model, and evaluates the total demand $\boldsymbol{D}$ attracted by each facility. The attracted demand evaluation $\boldsymbol{D}$ is sent to the congestion-sensitive subsystem, which compares it with capacity $\boldsymbol{X}$, evaluates new weights $\boldsymbol{W}$ and new accessibilities $\boldsymbol{A}$. Finally, the new accessibility evaluation is sent back again to the accessibility sensitive demand subsystem, and a new cycle is started.

(b) The smaller loop links the allocation and the congestion evaluation. This is a simple loop, nested within the main one, with the only purpose of updating the attractiveness vector $\boldsymbol{W}$, to be used in the allocation model, and the attracted demand vector $\boldsymbol{D}$, to be used in the evaluation of new congestion-sensitive weights.

A warning will conclude this section. The system has been described in pseudo-dynamic terms, but its use for a dynamic simulation is not suggested. Although the way it works is better understood in dynamic terms, the general model of figure 10 is more likely to give good steady-state solutions, rather than the actual transient behavior of an allocation system. This is mainly becuase of two reasons. The first is that most of the models used, especially the allocation (spatial-interaction) model, are equilibrium models. Usually the transient dynamic behavior of a spatial-interaction pattern does not fit a gravity model, and it takes some nonnegligible time to settle down (provided no further noise is introduced). The second reason is that usually real spatial-interaction systems have time lags. The reaction of demand to changes in accessibility and in congestion cannot be immediate, nor can demand be assumed to receive perfect information on all the changes in every location as soon as these changes take place.

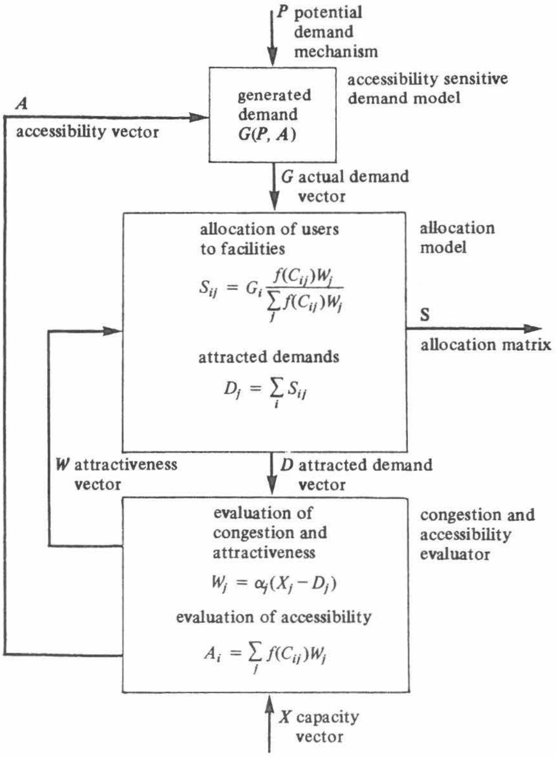
P potential demand mechanism
accessibility sensitive demand model
generated demand G(P, A)
A
accessibility vector
G actual demand vector
allocation of users to facilities
$S_{ij} = G_i \frac{f(C_{ij})W_j}{\sum_j f(C_{ij})W_j}$
allocation model
S
allocation matrix
attracted demands
$D_j = \sum_i S_{ij}$
W attractiveness vector
D attracted demand vector
evaluation of congestion and attractiveness
$W_j = \alpha_j(X_j - D_j)$
congestion and accessibility evaluator
evaluation of accessibility
$A_i = \sum_j f(C_{ij})W_j$
X capacity vector

Figure 10. The overall accessibility-congestion-sensitive allocation model.

In spite of this, a rough dynamic use of the model is possible when one is interested in long-term changes rather than in transient behavior. In other words, if information on system changes is sampled at time intervals at least as long as the average settling-down time, and if significant input changes can be reasonably assumed to take place only at the sampling times, then all that will be observed is a sequence of equilibrium states. This kind of analysis is closer to comparative statics, rather than dynamics. In any event, the construction of a satisfactory dynamic spatial-interaction framework is not an aim of this paper. [Some attempts in this direction can be found in Bertuglia and Leonardi (1979), and Wilson (1979).]

2.4 *Goals and tools for service location planning*

2.4.1 *Introduction.* Whereas the previous sections have been mainly devoted to the description of the system's behavior, section 2.4 will review the problem of controlling the location–allocation system in some optimal way. This means answering three questions.

1. Which goals are relevant for a service location planning decision?
2. Which variables can be used to control the systems, and under what constraints?
3. What is the general form of models corresponding to different goal-control variable combinations?

As already stated in the general introduction, no attempt will be made here to go deeply into such complicated problems as the foundations of welfare theory, public goods theory, equity theory, and so on. The existing location models have usually simple objective functions and constraints, and no large theoretical apparatus is needed to interpret and justify them. However, some common features in them are general enough to deserve a special discussion.

Most of these general problems have already been met in the introductory sections: the equity-efficiency trade-off problem, which is easily generalized to multiobjective conflicting problems; the location–allocation consistency problem; and the introduction of actual users' behavior in location models.

2.4.2 *The equity-efficiency trade-off problem.* A vast literature can be found on the conflict between equity and efficiency in location problems (see Alperovich, 1972; McGrew and Monroe, 1975; McAllister, 1976; Morrill and Symons, 1977; Bigman and ReVelle, 1978; 1979; and Lea, 1979). Although discussions on the exact definition of terms such as 'equity', 'welfare', and 'efficiency' tend to be endless, the problem behind these terms is quite simple. Apart from technical details, all measures of equity or welfare used in location problems are measures of nearness, ease of access, and fair distribution of service to users. This is true for the transport-cost minimization criterion used in most models (although some criticism can be raised against it as far as fair distribution is concerned), for the maximum coverage criterion used in many models for the location of emergency services, and for the consumer-surplus maximizing models used in locating urban activities with spatial interactions.

Maximization of access and equitable distribution of service capacity to the customers usually implies the opening of many dispersed small facilities. On the other hand, efficiency is usually measured in terms of the costs to be paid to establish and operate the facilities. Since both costs usually exhibit economies of scale, maximization of efficiency often implies concentration of service in a few large facilities. In loose terms, an optimal location model might be generally defined as a method to find a trade-off between these two conflicting goals. It is also worth noting that the existence of these two conflicting goals is the only reason location models have some meaning and usefulness. Locational decision taken according to only one goal would lead to ridiculous and unrealistic results.

From the technical standpoint, there are three ways to introduce equity-efficiency goals in location problems, giving rise to three different broad families of models:
1. Models where some user's benefit (cost) is maximized (minimized), subject to a budget constraint on total cost to establish and run the facilities. This is usually considered as the typical formulation for a public facility location problem, since public authorities, who are assumed to pay for the costs, are supposed not to use their budget for profit making. The above general definition of public facility location problems is found in ReVelle et al (1970), ReVelle and Rojeski (1970), Swain (1974), and Hansen and Kaufman (1976).
2. Models where some efficiency measure is maximized, subject to a constraint on the minimum users' benefit requirement. Typical examples are the so-called set-covering problems, which are also widely used to locate some public facilities (mainly emergency services). In its simplest form, a set-covering location problem minimizes the number of facilities to be located (the efficiency goal) subject to the requirement that each demand location has at least one facility within a given maximum range of travel time (the equity constraints). This formulation has been widely used for locating emergency services with deliveries from facilities to demand locations, such as fire stations and ambulance systems (Toregas et al, 1971; ReVelle et al, 1976; Plane and Hendrick, 1977).
3. Models where the objective function is the difference between a measure of users' benefit and public authority costs needed to establish and run facilities without any constraints. Since the two terms are usually measured in different units, at least one of them (usually the cost term) must be weighted with a scaling factor, reflecting the judgment of the decisionmaker (or decisionmakers) on their relative importance. Although this formulation is usually considered as more suited to private sector location problems, it has also been used to analyze trade-offs between benefits and costs in public facility location problems (possibly with a sensitivity analysis on the scaling factor). This 'bi-objective' approach is discussed in Bigman and ReVelle (1979) and Erlenkotter (1977).

Although these three formulations seem rooted on quite different public welfare conceptions, they share many common formal features, to such an extent that they can actually be reduced to the same mathematical form and solved with the same algorithms. If, for instance, a Lagrangian relaxation is introduced for the constraints of the first two types of models, both can be replaced by a model of the third type, the Lagrange multiplier acting as a scaling factor. This is even more sensible than the original formulations, since it is usually hard, for a public decisionmaker to assess a priori the values for a total budget or a minimum travel-time requirement, so that a trade-off sensitivity analysis has to be made anyway.

Other unifying properties can be found in the various ways of measuring travel cost and accessibility to the users. Although the minimum distance requirement seems very different from travel cost minimization, it has been shown (Church and ReVelle, 1976) that the minimum distance requirement is a special case of travel-cost minimization, provided an infinite cost is placed on displacements outside of the required range.

2.4.3 *The location–allocation consistency problem.* It has been stated in section 2 that location and allocation may be controlled by different decisionmakers. For the case of users' attracting systems, allocation is always left to the users, whereas location and size of facilities are decided by a public authority.

If the simplifying assumption is introduced that users always choose the nearest facility, the resulting location models belong to the class of plant-location problems found in operations research and management science literature. The recent development of tremendously powerful dual-based algorithms (Bilde and Krarup, 1977; Erlenkotter, 1978; van Roy and Erlenkotter, 1980; Wolsey, 1980) makes it hard to imagine any further improvement on theory and computation for these problems.

If, on the other hand, users are allocated to facilities according to a general spatial-interaction-based allocation model of the form proposed in section 2.3, then many new theoretical and computational problems arise.

The general structure of a (possibly dynamic) location model based on a spatial-interaction allocation rule is shown in figure 11.

The variables and symbols introduced in the diagram of figure 11 are defined as follows:

$\boldsymbol{P}$ is a vector whose components are the potential demands on each demand location i, $\boldsymbol{P} = [P_i]$;

$\Delta \boldsymbol{P}$ is a vector whose components are the unit time potential demand increments in each demand location i, $\Delta \boldsymbol{P} = [\Delta P_i]$;

$\boldsymbol{Y}$ is a vector whose components are the existing service capacities in each service location j, $\boldsymbol{Y} = [y_i]$;

$\boldsymbol{X}$ is a vector whose components are the new capacities in each service location j, after the location decision has taken place, $\boldsymbol{X} = [x_j]$;

S is the allocation matrix, that is, a matrix whose elements are the flows of users between each demand-facility location pair (i, j), $\mathrm{S} = [S_{ij}]$;

$\mathrm{A}(\boldsymbol{P}, \boldsymbol{X})$ is the allocation rule, that results from the general model shown in figure 10, that is, $\mathrm{A}(\cdot, \cdot)$ is a matrix function of the vectors of potential demand and service capacity, and the equation $\mathrm{S} = \mathrm{A}(\boldsymbol{P}, \boldsymbol{X})$ holds;

$V(\mathrm{S}, \boldsymbol{P}, \boldsymbol{X})$ is the public-authority-evaluated users' benefit, that is, $V(\cdot, \cdot, \cdot)$ is a scalar function of the allocation matrix, of the potential demand vector and of the capacity vector;

$C(\boldsymbol{Y}, \boldsymbol{X})$ is the cost paid by the public authority to change the capacity (by expansion or demolition), that is, $C(\cdot, \cdot)$ is a scalar function of the old and new capacity vectors;

α is a unit-time discount factor; and

$F(\boldsymbol{P}, \boldsymbol{Y})$ is the total discounted value for a decision process starting with potential demand $\boldsymbol{P}$ and capacity $\boldsymbol{Y}$.

The general model of figure 11 can be further specialized, depending on which one of the following assumptions holds:

(a) Both location and allocation are directly decided by the public authority; this often (but not necessarily) leads to a problem of the plant-location type. Of course,

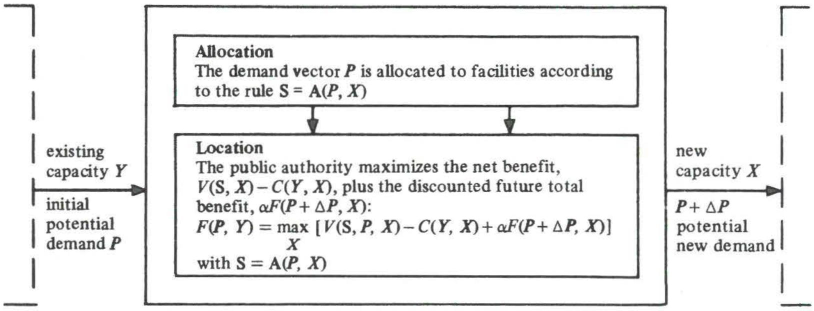
Allocation
The demand vector P is allocated to facilities according to the rule $\mathrm{S} = \mathrm{A}(P, X)$
Location
The public authority maximizes the net benefit, $V(\mathrm{S}, X) - C(Y, X)$, plus the discounted future total benefit, $\alpha F(P + \Delta P, X)$:
$F(P, Y) = \max_{X} [V(\mathrm{S}, P, X) - C(Y, X) + \alpha F(P + \Delta P, X)]$
with $\mathrm{S} = \mathrm{A}(P, X)$
existing capacity Y
initial potential demand P
new capacity X
$P + \Delta P$ potential new demand

Figure 11. The general form of a dynamic location-allocation model. The variables and symbols are defined in section 2.4.3.

in this case no spatial-interaction behavior takes place, since it is inhibited by the public authority-imposed behavior. (Many school allocation policies work like this, although they are often subject to serious criticisms.)

(b) Allocation results from unplanned users' behavior, that is, from a spatial-interaction model. The users' behavior model is known to the public authority, who takes it into account when deciding the location and size of facilities. However, the benefit evaluation of the public authority is not necessary related to some natural users' benefit function, nor does such a function need to exist. This is a case where possible conflicting problems may arise, both among users and between the users and the public authority. Of course, a spatial-interaction model is not suited to take conflicts into account, since it is an aggregate model in which the behavior of all users is averaged out. The microscopic approach seems better for this purpose (see the first approach discussed in section 2.3.2), since it can keep track of each user (or possibly of each homogenous group of users, if such a grouping can be defined). This approach will not be developed here, but it may well be one of the topics for future research on location models.

(c) As in assumption (b), allocation results from unplanned users' behavior, but, unlike assumption (b), some aggregate users' benefit function is assumed to exist. More precisely, a scalar function $W(\mathbf{S}, \boldsymbol{P}, \boldsymbol{X})$ can be defined, such that the problem

$$\underset{\mathbf{S}}{\text{maximize}}\ W(\mathbf{S}, \boldsymbol{P}, \boldsymbol{X}) \tag{6}$$

is solved by means of

$$\mathbf{S} = \mathbf{A}(\boldsymbol{P}, \boldsymbol{X})\,, \tag{7}$$

where $\mathbf{A}(\boldsymbol{P}, \boldsymbol{X})$ is the users' behavior-based allocation rule. Since users behave as if they were maximizing function (6), it is natural to call W the users benefit function. It is also natural, for a fair public decisionmaker, to use the same benefit function in the evaluation, that is, to make the definition

$$V(\mathbf{S}, \boldsymbol{P}, \boldsymbol{X}) = W(\mathbf{S}, \boldsymbol{P}, \boldsymbol{X})\,. \tag{8}$$

When assumptions (6), (7), and (8) hold, no conflict problem arises between the users and the public decisionmaker (PDM), that is, there is perfect consistency between location and allocation.

Assumptions (a), (b), and (c) give rise to the following three families of mathematical programming problems. (For simplicity, a static formulation is used; that is, $\alpha = 0$):

$$\underset{\mathbf{A},\boldsymbol{X}}{\text{maximize}}\{V[\mathbf{A}(\boldsymbol{P}, \boldsymbol{X}), \boldsymbol{P}, \boldsymbol{X}] - C(\boldsymbol{Y}, \boldsymbol{X})\}\,, \tag{9a}$$

that is, the PDM decides both on the allocation rule A and on the size of location of facilities $\boldsymbol{X}$;

$$\underset{\boldsymbol{X}}{\text{maximize}}\{V[\mathbf{A}(\boldsymbol{P}, \boldsymbol{X}), \boldsymbol{P}, \boldsymbol{X}] - C(\boldsymbol{Y}, \boldsymbol{X})\}\,, \tag{9b}$$

that is, the PDM decides on the size and location of facilities $\boldsymbol{X}$, with the allocation rule A given and depending on users' behavior; and

$$\underset{\boldsymbol{X}}{\text{maximize}}\{\underset{\mathbf{A}}{\text{maximize}}\ V[\mathbf{A}(\boldsymbol{P}, \boldsymbol{X}), \boldsymbol{P}, \boldsymbol{X}] - C(\boldsymbol{Y}, \boldsymbol{X})\}\,, \tag{9c}$$

that is, the PDM decides on the size and location of facilities $\boldsymbol{X}$, whereas the users decide on the allocation rule A. Moreover, both the PDM and the users agree on the way to evaluate benefits, since users choose, among all possible allocation rules, the one that maximizes V, whereas the PDM chooses, among all possible location patterns $\boldsymbol{X}$, the one that maximizes $V - C$.

2.4.4 *Spatial-interaction embedding location models.* A noteworthy property of problem (9c) is that, since the order with which the maximization operators are applied is immaterial, it can be rewritten as:

$$\underset{\mathbf{A}}{\text{maximize}}\ \underset{X}{\text{maximize}}\{V[\mathbf{A}(P,X),P,X]-C(Y,X)\}\,, \tag{10}$$

and this is identical to (9a). This means that, in the case of a perfect consistency between location and allocation goals the location problem is formally equivalent to a totally planned problem, although with a generally nonlinear objective function.

Let two special, but still fairly general, assumptions now be introduced:

(a) The cost function is separable, that is, the total cost can be expressed as the sum of individual costs for each facility location:

$$C(Y,X) = \sum_j h_j(x_j)\,, \tag{11}$$

where $h_j(\cdot)$ is a (usually concave) scalar function. The Y vector has been dropped, since it is constant, so that it is implicitly taken into account in the way the $h_j(\cdot)$ are built.

(b) The function $V(\mathbf{S},P,X)$ can be expressed as

$$V(\mathbf{S},P,X) = \Psi(\mathbf{S},P-G,X-D)\,, \tag{12}$$

where G is the vector of demands generated in each i, $G = \left[\sum_j S_{ij}\right]$, and D is the vector of demands attracted in each j, $D = \left[\sum_j S_{ij}\right]$.

The meaning of assumption (12) is that users, as well as the public decisionmaker, evaluate the allocation pattern $\mathbf{S}$, the unsatisfied demand $P-G$ (generated by the accessibility-sensitive mechanism), and the difference between capacity and attracted demand $X-D$ (generated by the congestion-sensitive mechanism). If the following variables are introduced:

$$S_{i0} = P-G_i = P_i-\sum_j S_{ij}\,,$$

that is, S_{i0} is the unsatisfied potential demand in demand location i; and

$$S_{0j} = x_j-D_j = x_j-\sum_i S_{ij}$$

that is, S_{0j} is the unused capacity in facility j; and the $\mathbf{S}$ matrix is augmented by one column $[S_{i0}]$ and one row $[S_{0j}]$ (the element S_{00} is not required), then an alternative representation of function ψ is:

$$\Psi(\mathbf{S}) \tag{13}$$

subject to

$$\sum_j S_{ij}+S_{i0} = P_i\,, \tag{14}$$

$$\sum_i S_{ij}+S_{0j} = x_j\,. \tag{15}$$

It is also tacitly assumed that $\Psi(\cdot)$ is undefined for negative values, so that explicit nonnegativity constraints are not needed.

If assumptions (a) and (b) are introduced into problem (10), the following mathematical programming problem is obtained:

$$\underset{\mathbf{S},X}{\text{maximize}}\ [\Psi(\mathbf{S})-\sum_j h_j(x_j)]\,,$$

subject to

$$\sum_j S_{ij} + S_{i0} = P_i \quad \text{and} \quad \sum_i S_{ij} + S_{0j} = x_j \, .$$

But the x_j variables can be eliminated by means of constraints (15), when expressed in terms of the S variables. The general form of the resulting location-allocation problems is:

$$\underset{\mathbf{S}}{\text{maximize}} \left[\Psi(\mathbf{S}) - \sum_j h_j \left(\sum_i S_{ij} + S_{0j}\right)\right] , \tag{16}$$

subject to

$$\sum_j S_{ij} + S_{i0} = P_i \, . \tag{17}$$

The above formulation includes most of the existing static facility location models. As a special case, if demand is inelastic (that is, $S_{i0} = 0$, $S_{0j} = 0$), Ψ is linear, $h_j(\cdot)$ are made up of a fixed charge plus a linear cost, it reduces to the standard plant-location problem (provided nonnegativity restrictions for **S** are introduced). If, however, Ψ is built in such a way that a spatial-interaction model is induced for the optimal **S**, then problem (16)–(17) includes most spatial-interaction based location-allocation models.

The most useful property of such a formulation is that the spatial-interaction model need not be introduced explicitly in the objective function or in the constraints. It is 'embedded' by the optimization, and it holds for the optimal point, although it is not required to hold in the whole of the feasible region. The above two devices, the embedding property and the widening of the feasible region, are the main reasons that models of the type given by problem (16)–(17) are computationally attractive and successful. The embedding approach had been introduced first by Coelho and Wilson (1976), and its close relationships with earlier works and problems were soon realized (Harris, 1964; Lakshmanan and Hansen, 1965; Huff, 1966; Leonardi, 1973). Among its recent developments, the most promising ones for further research seem to be the extention to more complex systems of spatially-interacting urban activities (Wilson, 1978; Wilson and Macgill, 1979; Leonardi, 1979a; Macgill and Wilson, 1979) with the possible inclusion of the transport network optimization (Boyce and Southworth, 1979; Boyce and LeBlanc, 1979).

There is, however, a general criticism to be made to most of the works referred to above. Although much theoretical insight has been gained by developing sound economical interpretations for the proposed spatial-interaction embedding objective functions (Neuburger, 1971; McFadden, 1974a; 1974b; Williams, 1977), most of the cost functions introduced in the models seem too simple and unrealistic; namely, they are usually simple linear functions, which, of course, cannot account for any economy-of-scale effect. In this respect, the traditional plant-location problems seem much more realistic, whether they introduce the scale effect by a nonsmooth fixed-charge cost function, thus making it a combinatorial programming problem, or they use a smooth concave cost function. The need is felt, therefore, to develop new models, which share both the realistic features of the spatial-interaction embedding functions, and the realistic features of the concave (possibly nonsmooth) cost functions. Some attempts have been made already in this direction, by introducing general concave functions, such as in Jacobsen and Kessel (1977), or by means of concave power functions, as in Leonardi (1978; 1979a), or fixed charges and combinatorial structures, as in Hodgson (1978), Beaumont (1979), and Leonardi (1980). This strand of research seems a most promising one, as it promotes reconciliation between the two 'schools' (the OR approach and the spatial-interaction approach), and it yields more useful models for real planning problems. It is, therefore, a leading topic for the second part of this paper, as well as for further research within the Public Facility Location Task.

2.4.5 *An overview of the relationships among static facility-location models.* It is worthwhile at this point to stop and to look back at what has been discussed, before introducing new topics. An overall picture of the links among assumptions and models is shown in figure 12. Basically, three broad families of location models have been found (the rounded boxes at the bottom of the diagram):

1. *The plant-location and related linear models.* Although these are borrowed from private-sector location problems, they have been shown to be useful for some public service location problems as well, such as, delivery systems and emergency services. Their use for some users-attracting systems (such as primary schools) has also been found, although the soundness of this approach is not warranted.

2. *The spatial-interaction embedding models.* These have been developed mainly in the fields of regional science and urban planning and are still waiting for further development. Although their beginning has been in retail location problems (Harris, 1964), they seem to be a general tool to model all users' attracting systems, and

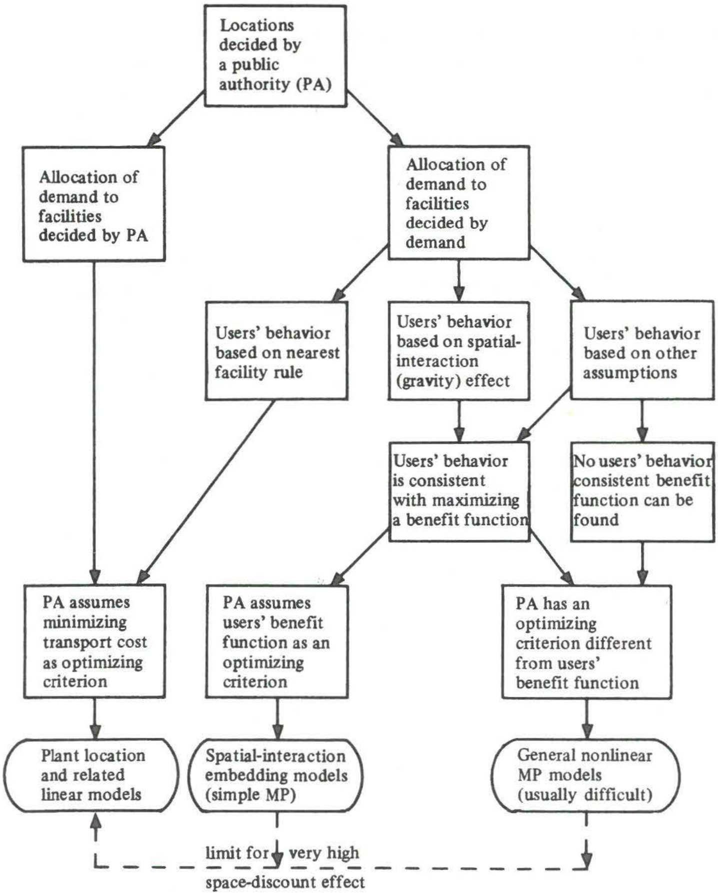
Locations decided by a public authority (PA)
Allocation of demand to facilities decided by PA
Allocation of demand to facilities decided by demand
Users' behavior based on nearest facility rule
Users' behavior based on spatial-interaction (gravity) effect
Users' behavior based on other assumptions
Users' behavior is consistent with maximizing a benefit function
No users' behavior consistent benefit function can be found
PA assumes minimizing transport cost as optimizing criterion
PA assumes users' benefit function as an optimizing criterion
PA has an optimizing criterion different from users' benefit function
Plant location and related linear models
Spatial-interaction embedding models (simple MP)
General nonlinear MP models (usually difficult)
limit for very high space-discount effect

Figure 12. A summary of the main relationships among different location models.

their use for public facility location has probably more justifications than their original use in the retail sector (since they are based on users' benefit maximizing, rather than on a producer's profit maximizing).

3. *The models which do not belong either to 1 or 2.* These models are a minority in the literature, and most of them were probably produced because of lack of insight into their economical justifications. There are, however, a few exceptions which deserve further attention, since they seem to pose some new meaningful and challenging issues for future research. Some examples are the work of Hanjoul (1980) on the disaggregate analysis of users' utility functions and preference orders; of Hansen et al (1979) on the introduction of pricing policies in locational decisions; of Hansen and Thisse (1980) on the introduction of voting procedures in locational decisions; and of Stahl (1980) on locational patterns arising from imperfect users' information.

Underlying these three broad families of models, a hierarchy of assumptions (the square boxes in figure 12) can be built, each path in the hierarchy leading to just one of the sets of models. The following five hierarchical levels can be seen. From top to bottom they are:

(1) An assumption on who decides location and size of facilities. (By the definition of a public facility location, this decisionmaker can only be a public authority.)

(2) An assumption on who decides demand allocation to facilities. (This is the very start of possible branchings, leading to different models.)

(3) An assumption on users' behavior in choosing facilities.

(4) An assumption on the existence of an overall users' benefit function.

(5) An assumption on the consistency between public authority goals and users' goals.

Various combinations of different answers to the questions posed in the assumptions above lead to different models, as is also shown in figure 12.

Another useful property for a unifying framework is revealed by the dotted-line at the bottom of figure 12. All the existing models seem to include the linear plant-location model as a special limiting case. This property holds true for the spatial-interaction embedding models (Leonardi, 1980), and it is rooted in an earlier well-known result (Evans, 1973). Some care, however, should be taken when stating that this always holds true for the third kind of models (the rounded box on the right-hand side of figure 12), although this can be shown for some meaningful cases (Hodgson, 1978; Hanjoul, 1980). Anyway, this limiting property has to be recommended as a criterion for choosing between good and bad models, since usually models which do not meet it are easily found to have some inconsistency in theory and formulation.

2.5 *Towards more general space–time structures*

2.5.1 *Introduction.* Although the problems of locating different interacting activities and of dynamic (multistage) optimization have been mentioned, the general framework developed so far is mainly a static single-service one. This reflects the state of the existing literature, which is mainly concerned with static single-service location problems. It is felt, however, that the future of location planning tools is in dynamic multiservice (or multiactivity) location models.

The reason for multiservice systems becomes apparent if, for instance, the assumption of home-based return trips is dropped, and more general multiple-destination trips are allowed for the users. Multiple trips link all services and activities together (including residence), and a public decision on how to optimally locate them has to take these links into account. The reason for a dynamical formulation becomes apparent when it is realized that services have to be located in a built environment, where the facilities are already existing. Most static location models are very unrealistic in this respect, as they seem to assume that everything has to be built from scratch, except demand locations and sizes. But the real urban management problem has to deal with existing physical stocks of facilities, which means that capacity expansion, demolition, and relocation, rather than simple-minded optimal location, are the meaningful decisions. Of course, a dynamic framework is required for such decisions, since the urban manager has to take into account both the past (the current demand and stock of facilities) and the future (demand changes and influence of changes in current stock on future demand-facility interaction).

The above stated problems cannot be treated with the same detail used for static facility location problems, since they need a substantial amount of new theoretical and empirical research, which has still to be carried out. Only an outline will be given here for some of the possible issues for future research programs, with no claim of exhaustiveness.

2.5.2 *The multiactivity location problem.* The multiactivity location problem dates back as far as Koopmans and Beckmann's (1957) paper, for a normative approach, and to Lowry's (1964) paper, for a descriptive approach. The simple, and yet powerful, Lowry framework has been given a mathematical-programming interpretation by Coehlo and Williams (1978), Wilson (1978), Leonardi (1979a), Macgill and Wilson (1979), and Shmulyan (1980), thus leading towards its possible use as an explicitly normative tool. Building on the above mentioned references, a possible multiactivity generalization of the model given by problem (16)–(17) is as follows:

$$\underset{\mathbf{S}}{\text{maximize}} \left[\Psi(\mathbf{S}) - \sum_{j,k} h_j^k \left(\sum_i S_{ij}^k + S_{0j}^k\right)\right] , \tag{18}$$

subject to

$$\sum_j S_{ij}^k + S_{i0}^k = P_i^k + \sum_r a_{rk} \sum_j S_{ji}^r , \tag{19}$$

where all variables and functions are defined in sections 2.4.3 and 2.4.4, except for the superscripts k and r, labeling different activities, and a_{rk}, the ratio of k-activity demand to r-activity demand. Sometimes (but not necessarily) the a_{rk} coefficients can be interpreted as transition probabilities. More generally, the $[a_{rk}]$ array introduces a linear technology tying all activities together; in this sense, constraints (19) can be interpreted as some kind of spatial breakdown of an input-output-like model. Indeed, if summation over i is taken on both sides of constraints (19), the following set of equations is obtained:

$$Q_k = Y_k + \sum_r Q_k a_{rk} , \tag{20}$$

where Q_k is the total actual (satisfied) demand for activity k, $Q_k = \sum_{ij} S_{ij}^k$; and Y_k is the difference between the potential demand for activity k originating in demand residences and the unsatisfied demand for activity k, $Y_k = \sum_i (P_i^k - S_{i0}^k)$. If the following vector and matrix notation is introduced:

$$\boldsymbol{Q} = [Q_k] , \qquad \boldsymbol{Y} = [Y_k] , \qquad \mathbf{A} = [a_{rk}] ,$$

then equations (20) become

$$\boldsymbol{Q} = \boldsymbol{Y} + \boldsymbol{Q}\mathbf{A} , \tag{21}$$

and its explicit solution, for $\boldsymbol{Y}$ given, is:

$$\boldsymbol{Q} = \boldsymbol{Y}(\mathbf{I} - \mathbf{A})^{-1} . \tag{22}$$

The resemblance of equation (22) to the general solution of an input–output model thus becomes evident. But the main interesting feature of equation (22) is that, unlike in the usual input–output case, the total demand vector *Y is not given*. Because of its definition, it depends on unsatisfied demand, which in turn depends on the accessibility-sensitive demand generation mechanism. This means that the total demand vector *is affected* by the spatial-interaction process. This is a warning for every top-to-bottom resource allocation procedure: feedbacks come from the bottom, and different locational policies may require changes in higher-level resource allocation policies.

2.5.3 *The dynamic-location problem*. The dynamic-location problem was first introduced in Manne (1967), in a mainly private-sector oriented fashion. It was further extensively reviewed by Scott (1971), Sheppard (1974), and Erlenkotter (1979). A new operational approach for solving dynamic-location problems has recently been developed by van Roy and Erlenkotter (1980). The trouble is, that all these works refer to linear objective functions, and no spatial-interaction behavior is taken into account. Moreover, they are mainly oriented towards capacity expansion for growing demand in a finite time horizon [although this assumption has been relaxed to some extent recently, see van Roy and Erlenkotter (1980)]. On the other hand, in most developed countries urban managers are faced with problems of fast fluctuating (sometimes declining) demand, the fluctuations taking place over time, space (because of migrations and residential mobility), and kind of services. (For instance, although total high school demand is usually growing everywhere, specializations required from the labor market may change over time.) At the same time, the average life of urban stocks is usually very long, and decisions on expansion, demolition, and relocation are forced to be slower than the demand-changing process.

If a rephrasing of the dynamic location problems in terms of classical operations research model forms is attempted, one might say that, whereas the capacity-expansion oriented approach raises a mainly job-scheduling problem, the urban stocks management approach raises mainly an inventory problem. This difference leads to two very different ways of defining an optimal policy, as well as two very different planning philosophies:

1. For finite-horizon capacity expansion, an optimal policy is a project; that is, a list of facilities to be opened and closed at each time period.
2. For unbounded-horizon urban stock management, an optimal policy is an inventory rule; that is, a rule which, given the current knowledge of actual demand, its possible forecasts, and the current stock levels in each location, determines the best way these stock levels should be changed.

Thus, whereas in approach 1 the main concern will be how much, where, and when to build, approach 2 will lead to problems such as uncertainty over future demand, adaptive updating mechanisms, and monitoring indicators for changes in the system requiring new decisions. The modeling of such systems will not be attempted here, but it is one of the expected results of future research.

The second part of the paper will describe mathematical models for static, single-service, facility location problems and their possible extensions and improvements.

References

Alperovich G, 1972 "Welfare criteria and models for locating public facilities" discussion paper 19, School of Architecture and Urban Planning, Princeton University, Princeton, NJ

Balinski M L, 1961 "Fixed cost transportation problems" *Naval Research Logistics Quarterly* 8 41–54

Beaumont J R, 1979 "Some issues in the application of mathematical programming in human geography" WP-256, School of Geography, University of Leeds, Leeds, England

Bertuglia C S, Leonardi G, 1979 "Dynamic models for spatial interaction" *Sistemi Urbani* **1**(2) 3-25

Bigman D, ReVelle C, 1978 "The theory of welfare considerations in public facility location problems" *Geographical Analysis* **10** 229-240

Bigman D, ReVelle C, 1979 "An operational approach to welfare considerations in applied public facility location models" *Environment and Planning A* **11** 83-95

Bilde O, Krarup J, 1977 "Sharp lower bounds and efficient algorithms for the simple plant location problem" *Annals of Discrete Mathematics* **1** 79-97

Boyce D E, LeBlanc L J, 1979 "Synthesizing network equilibrium and design models for location of urban activities" paper presented at the tenth International Symposium on Mathematical Programming, Montreal, Canada; available from Department of Civil Engineering, University of Illinois at Urbana-Champaign, Urbana, Ill.

Boyce D E, Southworth F, 1979 "Quasi-dynamic urban-location models with endogenously determined travel costs" *Environment and Planning A* **11** 575-584

Church R L, ReVelle C S, 1976 "Theoretical and computational links between the *p*-median, location set-covering, and the maximal covering location problem" *Geographical Analysis* **8** 406-415

Coelho J D, Williams H C W L, 1978 "On the design of land use plans through locational surplus maximization" *Papers of the Regional Science Association* **40** 71-85

Coelho J D, Wilson A G, 1976 "The optimum location and size of shopping centers" *Regional Studies* **10** 413-421

Coelho J D, Wilson A G, 1977 "An equivalence theorem to integrate entropy maximizing submodels within overall mathematical programming frameworks" *Geographical Analysis* **9** 160-173

Davidson K B, 1977 "Accessibility in transport/land-use modeling and assessment" *Environment and Planning A* **9** 1401-1410

Efroymson M A, Ray T L, 1966 "A branch-bound algorithm for plant location" *Operations Research* **14** 361-368

Erlenkotter D, 1977 "Facility location with price-sensitive demands: private, public, and quasi-public" *Management Science* **24** 378-386

Erlenkotter D, 1978 "A dual-based procedure for uncapacitated facility location" *Operations Research* **26** 992-1009

Erlenkotter D, 1979 "A comparative study of approaches to dynamic location problems" WP-292, Western Management Science Institute, University of California, Los Angeles, Calif.

Evans S P, 1973 "A relationship between the gravity model for trip distribution and transportation problems in linear programming" *Transportation Research* **7** 39-61

Evans S P, 1976 "Some models for combining the trip distribution and traffic assignment stages in the transport planning process" in *Traffic Equilibrium Methods* Eds M Beckmann, H P Künzi (Springer, Berlin) pp. 201

Hanjoul P, 1980 "Facility location under particular allocation rules" paper presented at the Workshop on Location and Distribution Management, European Institute for Advanced Studies in Management, Brussels, Belgium; available from Unité de Science et de Programmation Urbaines et Régionales, Université Catholique de Louvain, Louvain-La-Heuve, Belgium

Hansen P, Kaufman L, 1976 "Public facilities location under an investment constraint" in *Operational Research '75* Ed. K B Haley (North-Holland, Amsterdam)

Hansen P, Thisse J F, 1980 "Condorcet, Weber and Rawls locations" research program paper 4, Unite de Science et de Programmation Urbaines et Régionales, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Hansen P, Thisse J F, Hanjoul P, 1979 "Simple plant location under uniform delivery pricing" paper presented at the Third European Congress on Operations Research, Amsterdam; available from Unité de Science et de Programmation Urbaines et Régionales, Université Catholique de Louvain, Louvain-La-Heuve, Belgium; also forthcoming in *Locational Analysis of Public Facilities* Eds J F Thisse, H G Zoller in the series *Studies in Mathematical and Managerial Economics* (North-Holland, Amsterdam)

Hansen W G, 1959 "How accessibility shapes land use" *Journal of the American Institute of Planners* **25** 73-76

Harris B, 1964 "A model of locational equilibrium for retail trade" mimeograph, Penn-Jersey Transportation Study, Philadelphia, Pa

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

Hodgson M J, 1978 "Towards more realistic allocation in location-allocation models: an interaction approach" *Environment and Planning A* **10** 1273-1285

Huff D L, 1964 "Defining and estimating a trading area" *Journal of Marketing* **28** 37-48

Huff D L, 1966 "A programmed solution for approximating an optimum retail location" *Land Economics* **42** 293-303

Ingram D R, 1971 "The concept of accessibility: a search for an operational form" *Regional Studies* **5** 101-107

Jacobsen S K, Kessel O, 1977 "The concept of entropy in OR-modelling" working paper, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby, Denmark

Koopmans T C, Beckmann M J, 1957 "Assignment problems and the location of economic activities" *Econometrica* **25** 53-76

Lakshmanan T R, Hansen W G, 1965 "A retail market potential model" *Journal of the American Institute of Planners* **31** 134-143

Lea A, 1979 "Welfare theory, public goods, and public facility location" *Geographical Analysis* **11** 218-239

Leonardi G, 1973 "Localizzazione ottimale dei servizi urbani (Optimal location of urban services)" *Ricerca Operativa* **12** 15-43

Leonardi G, 1975 "Un nuovo algoritmo per il problema della localizzazione ottimale dei servizi urbani (A new algorithm for the optimal location of urban services)" *Proceedings of the Congress of the Italian Operations Research Association* (AIRO, Milano) pp 121-132

Leonardi G, 1976 "Alcune considerazioni teoriche e sperimentali sulla relazione tra accessibilità e affolamento nei problemi localizzativi (Some considerations on theory and practice of the relationship between accessibility and congestion in location problems)" in *Teoria dei Sistemi ed Economia* Eds S Lombardiui, A Ruberti (Il Mulino, Bologna) pp 271-290

Leonardi G, 1978 "Optimum facility location by accessibility maximizing" *Environment and Planning A* **10** 1287-1305

Leonardi G, 1979a "Some mathematical programming ideas within a generalized spatial interaction and activity framework" in *London Papers in Regional Science 10. Developments in Urban and Regional Analysis* Ed. M J Breheny (Pion, London) pp 28-47

Leonardi G, 1979b "Introduzione alla teoria dell' accessibilità (Introduction to accessibility theory)" *Sistemi Urbani* **1** 65-88

Leonardi G, 1980 "On the formal equivalence of some simple facility location models" WP-80-21, International Institute for Applied Systems Analysis, Laxenburg, Austria; paper also presented at the Workshop on Location and Distribution Management, at the European Institute for Advanced Studies in Management, Brussels, Belgium

Lowry I S, 1964 "A model of metropolis" RM-4035-RC, Rand Corporation, Santa Monica, Calif.

Macgill S M, Wilson A G, 1979 "Equivalences and similarities between some alternative urban and regional models" *Sistemi Urbani* **1** 1-40

Manne A S (Ed.), 1967 *Investment for Capacity Expansion: Size, Location, and Time-phasing* (MIT Press, Cambridge, Mass)

McAllister D M, 1976 "Equity and efficiency in public facility location" *Geographical Analysis* **8** 47-63

McFadden D, 1974a "Conditional logit analysis of qualitative choice behavior" in *Frontiers in Econometrics* Ed. P Zarembka (Academic Press, New York) pp 105-142

McFadden D, 1974b "The mea. irement of urban travel demand" *Journal of Public Economics* **3** 303-328

McGrew J C, Monroe C B, 1975 "Efficiency, equity, and multiple facility location" *Proceedings of the Association of American Geographers* **7** 142-146

Morrill R L, Symons J, 1977 "Efficiency and equity aspects of optimum location" *Geographical Analysis* **9** 215-225

Neuburger H L I, 1971 "User benefit in the evaluation of transport and land use plans" *Journal of Transport Economics and Policy* **5** 52-75

Plane D R, Hendrick T E, 1977 "Mathematical programming and the location of fire companies for the Denver Fire Department" *Operations Research* **25** 563-578

ReVelle C S, Marks D, Liebman J C, 1970 "An analysis of private and public sector location models" *Management Science* **16** 692-702

ReVelle C, Rojeski P, 1970 "Central facilities location under an investment constraint" *Geographical Analysis* **2** 343-360

ReVelle C, Toregas C, Falkson L, 1976 "Applications of the location set-covering problem" *Geographical Analysis* **8** 65-76

Scott A J, 1971 *Combinatorial Programming, Spatial Analysis and Planning* (Methuen, Andover, Hants)

Sheppard E S, 1974 "A conceptual framework for dynamic location-allocation analysis" *Environment and Planning A* 6 547-564

Sheppard E S, 1979 "Geographic potentials" *Annals of the Association of American Geographers* 69 438-447

Shmulyan B L, 1980 "Spatial modeling of urban systems: entropy approach" CP-80-13, International Institute for Applied Systems Analysis, Laxenburg, Austria

Smith T E, 1976 "Spatial discounting and the gravity hypothesis" *Regional Science and Urban Economics* 6 331-356

Stahl K, 1980 "Location of markets under imperfect consumer information" paper presented at the Workshop on Location and Distribution Management, European Institute for Advanced Studies in Management, Brussels, Belgium

Steward J Q, 1948 "Concerning 'social physics'" *Scientific American* 178 20-23

Swain R W, 1974 "A parametrix decomposition approach for the solution of uncapacitated location problems" *Management Science* 21 189-198

Toregas C, Swain R, ReVelle C, Bergman L, 1971 "The location of emergency service facilities" *Operations Research* 19 1363-1373

van Roy T J, Erlenkotter D, 1980 "A dual-based procedure for dynamic facility location" WP-80-31, International Institute for Applied Systems Analysis, Laxenburg, Austria

Vickerman R W, 1974 "Accessibility, attraction, and potential: a review of some concepts and their use in determining mobility" *Environment and Planning A* 6 675-691

Weibull J W, 1976 "An axiomatic approach to the measurement of accessibility" *Regional Science and Urban Economics* 6 357-379

Williams H C W L, 1977 "On the formation of travel demand models and economic evaluation measures of user benefit" *Environment and Planning A* 9 285-344

Williams H C W L, Senior M L, 1978 "Accessibility, spatial interaction and the evaluation of land use transportation plans" in *Spatial Interaction Theory and Planning Models* Eds A Karlqvist, L Lundqvist, F Snickars, J W Weibull (North-Holland, Amsterdam)

Wilson A G, 1970 *Entropy in Urban and Regional Modelling* (Pion, London)

Wilson A G, 1974 *Urban and Regional Models in Geography and Planning* (John Wiley, Chichester, Sussex)

Wilson A G, 1976 "Retailers' profit and consumers' welfare in a spatial interaction shopping model" in *London Papers in Regional Science 6. Theory and Practice in Regional Science* Ed. I Masser (Pion, London) pp 42-59

Wilson A G, 1978 "Spatial interaction and settlement structure: towards an explicit central place theory" in *Spatial Interaction Theory and Planning Models* Eds A Karlqvist, L Lundqvist, F Snickars, J W Weibull (North-Holland, Amsterdam)

Wilson A G, 1979 "Some new sources of instability and oscillation in dynamic models of shopping centers and other urban structures" WP-267, School of Geography, University of Leeds, Leeds, England

Wilson A G, Macgill S M, 1979 "A systems analytical framework for comprehensive urban and regional model building" *Geographia Polonica* 42 9-25

Wolsey L A, 1980 "Fundamental properties of certain discrete location problems" paper presented at the Workshop on Location and Distribution Management, at the European Institute for Advanced Studies in Management, Brussels, Belgium; forthcoming in *Locational Analysis of Public Facilities* Eds J F Thisse, H G Zoller in the series *Studies in Mathematical and Managerial Economics* (North-Holland, Amsterdam)

Environment and Planning A, 1981, volume 13, pages 1085-1108

A unifying framework for public facility location problems—part 2: Some new models and extensions

G Leonardi
International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria
Received 19 June 1980

Abstract. This second part of the paper describes mathematical models for static, single-service, facility location problems and their possible extensions and improvements. The first part that appeared in the last issue (pages 1001-1028) gave a nontechnical description of the proposed general framework for analyzing location problems.

3 Some unifying proposals for static facility location models

3.1 *Introduction*

This section explores the properties and the possible solution methods for problem (16)-(17):

$$\underset{\mathbf{S}}{\text{maximize}} \left[\Psi(\mathbf{S}) - \sum_j h_j \quad \sum_i S_{ij} + S_{0j} \right] , \tag{16}$$

subject to

$$\sum_j S_{ij} + S_{i0} = P_i , \tag{17}$$

in the special case when a spatial-interaction model given in equation (1) holds, that is, $S_{ij} = G_i f(C_{ij}) W_j / \sum_j f(C_{ij}) W_j$, and the space-discount function is a negative exponential. It has been shown by Neuburger (1971) that the spatial-interaction embedding function for these models has the general form

$$\Psi(S) = -\frac{1}{\beta} \sum_{i,j} S_{ij} \log S_{ij} - \sum_{ij} S_{ij} C_{ij} , \tag{23}$$

where β is the space discount factor. Neuburger also shows that, if the spatial interaction model is looked at as a demand function (transport costs C_{ij} being the prices), then function (23) can be interpreted as the consumers' surplus associated with the allocation matrix $[S_{ij}]$ and the travel cost matrix $[C_{ij}]$. The function defined by equation (23) is therefore the most natural measure of users' benefit, consistent with the assumed spatial-interaction behavior, and the best suited evaluation criterion for a public decisionmaker. Function (23) is also related to Wilson's entropy function (Wilson, 1970), although, as Neuburger points out, the analogy is only formal and no deep meaning should be sought in it.

It will be noticed in equation (23) that, if $\beta \to \infty$, then $\Psi(S)$ reduces to a simple linear total transport cost term. Therefore, linear problems of the plant-location type are included in equation (23) as special cases.

The aim of the next sections is both to review the main existing formulations of the spatial-interaction embedding location problems, and to propose new possible models.

The exposition will be in order of increasing complexity, starting with the simplest case of inelastic-demand linear-budget constrained models, and gradually introducing elastic demand, existing stock, and nonlinear cost functions. All the problems are formulated in the first standard form discussed in section 2.4.2 of part 1 of the paper (pages 1016-1017), that is, maximizing users' benefit subject to a constraint on total budget.

3.2 *Linear budget constraints*

3.2.1 *The fixed demand case.* If a Neuburger consumer surplus function of the form (23) is assumed as a measure of users' benefit, and if the cost for establishing and running the facilities are linear functions of the form:

$$h_j(x_j) = a_j x_j + b_j \ ,$$

where

$h_j(x_j)$ is the total cost for a facility of size (capacity) x_j in location j,

a_j is the cost per unit size in location j, and

b_j is the fixed (size-independent) cost to be paid for opening a facility in j.

Then the general location model with fixed total demand assumes the following form:

$$\underset{\mathbf{S}, L, \mathbf{X}}{\text{minimize}} \ \frac{1}{\beta} \sum_i \sum_{j \in L} S_{ij} \log S_{ij} + \sum_i \sum_{j \in L} S_{ij} C_{ij} \ ,$$

subject to

$$\sum_{j \in L} S_{ij} = P_i \ , \qquad \sum_i S_{ij} \leqslant x_j \ , \qquad j \in L \ ,$$

$$\sum_{j \in L} (a_j x_j + b_j) \leqslant B \ , \qquad x_j \geqslant 0 \ , \qquad j \in L \ ,$$

(the c‿nstraints $S_{ij} \geqslant 0$ are not needed, since the objective function is not defined for negative values), where

i labels the demand locations, which are given;

j labels the locations of facilities;

L is the subset of chosen facility locations, among all possible given locations; the subset L is not given, but it has to be determined by optimization;

S_{ij} is the flow of users from demand location i to facility j, for $j \in L$ (here only users-attracting systems are considered);

P_i is the total demand in location i;

C_{ij} is the cost of a trip from demand location i to facility in location j;

x_j is the size of the facility in location j, for $j \in L$; and

B is the total budget available for establishing and running all the facilities.

The problem of choosing the subset L induces combinatorial features in the above mathematical program. If, as a special case, $\beta \rightarrow \infty$, the first term in the objective function disappears, and the above problem becomes:

$$\underset{\mathbf{S}, L, \mathbf{X}}{\text{minimize}} \sum_i \sum_{j \in L} S_{ij} C_{ij} \ ,$$

subject to

$$\sum_{j \in L} S_{ij} = P_i \ , \qquad \sum_i S_{ij} \leqslant x_j \ , \qquad j \in L \ ;$$

$$\sum_{j \in L} (a_j x_j + b_j) \leqslant B \ , \qquad x_j \geqslant 0 \ , \qquad S_{ij} \geqslant 0 \ , \qquad j \in L \ .$$

(Now the constraints $S_{ij} \geqslant 0$ are needed.)

Since there is no need to maintain unused capacity, when the budget constraint is binding (the most sensible case) it will also be:

$$x_j = \sum_i S_{ij} \ , \qquad j \in L \ .$$

That is, the total capacity of the facility in location j is set equal to the total demand attracted in j. The variables x_j can thus be eliminated, and the resulting location problem, containing the variables S_{ij} and the subset L only, is as follows:

$$\underset{\mathbf{S},L}{\text{minimize}} \sum_i \sum_{j \in L} S_{ij} C_{ij} ,$$

subject to

$$\sum_{j \in L} S_{ij} = P_i , \qquad \sum_{j \in L} (a_j x_j + b_j) = B , \qquad S_{ij} \geqslant 0 .$$

The above problem is the standard budget-constrained location problem found in the operations research literature (ReVelle et al, 1970; ReVelle and Rojeski, 1970; Hansen and Kaufman, 1976). It has thus been shown how, when the spatial discount rate β tends to infinity, the spatial-interaction-based location model tends to the plant-location model, and the induced limiting form in users' behavior follows the nearest-facility allocation rule.

In the most meaningful cases, however, β will not take the value infinity, and the plant-location models cannot be used. The resulting nonlinear combinatorial problems are discussed in Leonardi (1980), and results are forthcoming on efficient algorithms to solve these problems and make sensitivity analysis on the space discount rate and the budget level.

However, without going into the technical details of the algorithmic problems (which are outside of the scope of this paper), much insight can be gained in understanding the structure of the spatial-interaction-based location model, if its combinatorial part (that is, the set L) is held constant. If L is given, it can be dropped from the list of control variables, and the problem can be rewritten as

$$\underset{\mathbf{S},x}{\text{minimize}} \frac{1}{\beta} \sum_{i,j} S_{ij} \log S_{ij} + \sum_{i,j} S_{ij} C_{ij} , \tag{24}$$

subject to

$$\sum_j S_{ij} = P_i , \tag{25}$$

$$\sum_i S_{ij} \leqslant x_j , \tag{26}$$

$$\sum_j a_j x_j \leqslant R , \tag{27}$$

$$x_j \geqslant 0 , \tag{28}$$

where R, $R = B - \sum_{j \in L} b_j$, is the remaining budget after the fixed costs for opening the facilities have been deduced; with no loss of generality, R will also be called 'the budget' from now on.

Problem (24)-(28) is a convex programming problem with linear constraints. [It can be easily shown that the objective function (24) is convex.] It therefore has a unique optimal solution, and the Kuhn-Tucker first order conditions are necessary and sufficient to identify it. The special structure of the objective function, together with the simplicity of the constraints, makes it possible to find explicit closed-form solutions for the flow variables S_{ij} and for the facility sizes x_j, and to analyze the sensitivity to changes on the budget level R. Three cases can be given:

(a) The budget is more than is needed, so that constraint (27) is not binding.

(b) The budget is scarce, but sufficient to satisfy all the required capacity, so that constraint (27) is binding.

(c) The budget is not enough to satisfy all the required capacity, so that no solution to inequalities (26) can be found.

In the first case constraint (27) can be dropped, and also constraints (26) and (28) lose meaning. The optimal values for the flow variables S_{ij} are the solution to the following mathematical programming problem.

$$\text{minimize } \frac{1}{\beta}\sum_{i,j} S_{ij}\log S_{ij} + \sum_{ij} S_{ij}C_{ij}\,, \tag{29}$$

subject to

$$\sum_{j} S_{ij} = P_i\,. \tag{30}$$

By means of the standard Lagrange-multipliers method, one can easily obtain for the optimal S_{ij}:

$$\overline{S}_{ij} = P_i\frac{f_{ij}}{\sum\limits_{j} f_{ij}}\,, \tag{31}$$

where

$$f_{ij} = \exp(-\beta C_{ij})\,. \tag{32}$$

The conditions can now be stated under which expression (31) is actually the solution to problem (24)-(28). The total size of the facility in j must not be less than the total demand attracted in j, so that its minimum feasible value is

$$\overline{x}_j = \sum_{i}\overline{S}_{ij} = \sum_{i} P_i\frac{f_{ij}}{\sum\limits_{j} f_{ij}}\,. \tag{33}$$

From expression (33) it follows that the total cost of the location plan cannot be less than $\overline{R}$, where

$$\overline{R} = \sum_{j} a_j\overline{x}_j\,. \tag{34}$$

Thus, if $R \geqslant \overline{R}$, $\overline{x}_j$ as computed from expression (33) is an optimal solution to problem (24)-(28); and if, $R < \overline{R}$, the budget constraint becomes binding and can no longer be neglected.

The second case, where $R < \overline{R}$, applies when the budget satisfies the inequalities

$$\left(\sum_{i} P_i\right)\left(\min_{j} a_j\right) \leqslant R < \overline{R}\,, \tag{35}$$

where $\overline{R}$ is given by expression (34). The meaning of the left-hand side of condition (35) is as follows: the lowest cost solution is to locate just one big facility in the cheapest location, regardless of travel costs for the users. It is required that the budget be not less than the cost of the lowest cost solution.

Because of condition (35), constraint (27) is now binding, so that it is always a strict equality. This will be true for constraints (26) as well, since a slack capacity in the facilities could only be afforded if one had an excess budget, or if $R > \overline{R}$, which contradicts condition (35).

The x_j variables can be eliminated, since the size of each facility is equal to the number of users it attracts. Problem (24)-(28) reduces to

$$\text{minimize } \frac{1}{\beta}\sum_{i,j} S_{ij}\log S_{ij} + \sum_{ij} S_{ij}C_{ij}\,, \tag{36}$$

subject to

$$\sum_{j} S_{ij} = P_i\,, \tag{37}$$

$$\sum_{j} a_j\sum_{i} S_{ij} = R\,, \tag{38}$$

and use of the Lagrange-multiplier technique yields for the general solution

$$S_{ij} = P\frac{f_{ij}\exp(-\lambda a_j)}{\sum\limits_j f_{ij}\exp(-\lambda a_j)} , \tag{39}$$

where λ is the Lagrange multiplier for the budget constraint (38).

A comparison of expressions (39) and (31) shows that expression (31) reduces to expression (39) if $\lambda = 0$. The term $\exp(-\lambda a_j)$ can thus be considered as a special attraction weight for each location, attractiveness decreasing with unit cost in the location. Each location has a different attraction weight and only when the budget constraint is not binding, that is when $\lambda = 0$, will all locations have equal weights. One remark is perhaps needed on these attractiveness weights. From the way they are built, one might argue that prices b_j are charged to the users. But no pricing policy is assumed here. Constraints (26) have been dropped by reason of algebraic manipulation only, but they are *still acting*. What users perceive is the capacity offered in j, x_j, *and not* the unit cost b_j, which is paid by the public decisionmaker. Thus, in spite of how it looks, expression (39) arises from a doubly constrained spatial-interaction behavior.

It is useful to analyze the size of the facilities as a function of λ, in order to find the sensitivity of the solution to changes in the available budget. Let us make the following definition:

$$x_j(\lambda) = \sum_i S_{ij} = \sum_i P_i \frac{f_{ij}\exp(-\lambda a_j)}{\sum\limits_j f_{ij}\exp(-\lambda a_j)} , \tag{40}$$

It can be shown that the derivative of expression (40) is given by

$$x_j' = \sum_i \bar{a}_i S_{ij} - a_j x_j , \tag{41}$$

where S_{ij} and x_j are given by expressions (39) and (40), and $\bar{a}_i$ is defined by

$$\bar{a}_i = \frac{\sum\limits_j S_{ij} a_j}{P_i} ,$$

that is, the average unit cost weighted by the flows coming from location i.

Derivatives (41) vanish if

$$a_j = \frac{\sum\limits_i \bar{a}_i S_{ij}}{\sum\limits_i S_{ij}} , \tag{42}$$

and, if equation (42) has a solution, convexity of expression (36) assures it is an optimal one for problem (36)–(38).

Other easily found results are

$$x_j(0) = \bar{x}_j , \tag{43}$$

$$x_j(\infty) = \begin{cases} \sum\limits_i P_i , & \text{if} \quad a_j = \min\limits_k a_k , \\ 0 , & \text{if} \quad a_j > \min\limits_k a_k , \end{cases} \tag{44}$$

$$x_j'(0) = \sum_i P_i \frac{f_{ij}}{\sum\limits_j f_{ij}} \left(\frac{\sum\limits_j f_{ij} a_j}{\sum\limits_j f_{ij}} - a_j \right) . \tag{45}$$

The above quantities can be used to classify locations into three main groups:

1. If $x_j'(0) > 0$ and $a_j = \min\limits_k a_k$, the graph of the function $x_j(\lambda)$ has the shape shown in figure 13. That is, the size of the facility in the location with minimum cost is increasing with λ, and it attracts all the demand in the limit when $\lambda \to \infty$ and R approaches the value of the left-hand side of constraint (35). In other words, when resources become scarce, the location cost becomes the main criterion to decide on size and location, and the optimal solution tends to concentrate in one single location (namely, the cheapest one).

2. If $x_j'(0) > 0$ and $a_j > \min\limits_k a_k$, the graph of the function $x_j(\lambda)$ has the shape shown in figure 14. Equation (42) has a finite solution λ^* so that the size x_j rises to a maximum x_j^* for $0 < \lambda \leqslant \lambda^*$, and then falls again for $\lambda > \lambda^*$. In other words, for this kind of location there is a threshold value for the budget, above which it is convenient to increase the size of the facility there, and below which the facility must be abandoned in favor of cheaper locations.

3. If $x_j'(0) < 0$ and $a_j > \min\limits_k a_k$, the graph of the function $x_j(\lambda)$ is shaped as in figure 15. Equation (42) has no finite solution in this case, so that the size x_j decreases monotonically with λ. This is clearly the worst kind of location, as far as scarce resources are concerned. The size of the facility there decreases for any decrease in the budget.

The previous analysis, simple as it is, reveals two facts:

(a) Even with linear costs (that is, with no scale economies) concentration effects can be caused by low budgets. This is mainly a result of the doubly constrained nature of the model, and in particular of constraint (26), requiring demand not exceeding capacity for each facility.

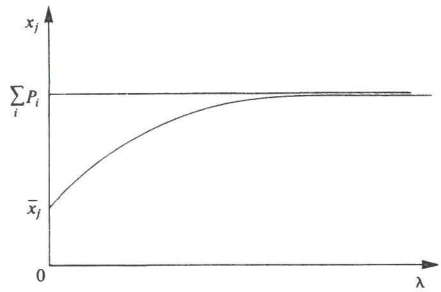

Figure 13. Size of facility as a function of the budget multiplier. Case 1: cheapest location.

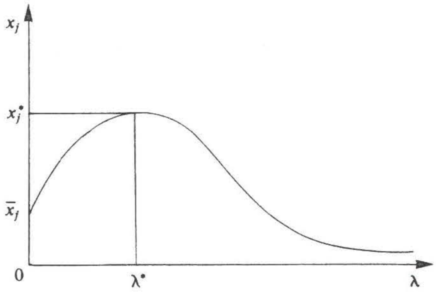

Figure 14. Size of facility as a function of the budget multiplier. Case 2: intermediate-cost location.

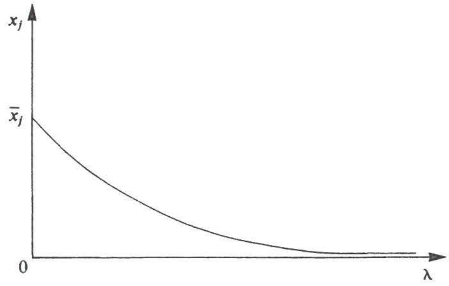

Figure 15. Size of facility as a function of the budget multiplier. Case 3: high-cost location.

(b) There is nothing like a fixed hierarchy among the sizes in each location. Changes in the total budget can turn the distribution of facilities upside down, so that what seemed to be a good location may be abandoned and what seemed to be a bad location may be chosen.

Parametric formulas (40) can be used quite easily to get actual numerical solutions, and these can be used to build up the total cost function:

$$T(\lambda) = \sum_j a_j x_j(\lambda) , \tag{46}$$

which can be easily shown to be monotonically nonincreasing for $\lambda \geqslant 0$ and such that $T(0) = \overline{R}$, and $T(\infty) = \left(\min_k a_k\right) \sum_i P_i$, where $\overline{R}$ is defined by expression (34). [The graph of $T(\lambda)$ is shown in figure 16.] Provided the actual budget R satisfies inequality (35), the optimal value of the multiplier λ is the root of

$$T(\lambda) = R , \tag{47}$$

which can be quite easily solved by numerical methods.

In the third case one has:

$$0 \leqslant R < \left(\min_k a_k\right) \sum_i P_i , \tag{48}$$

that is, the budget is not even enough to concentrate all the demand in the cheapest zone. As can be seen in figure 16 there is no real solution to equation (47) in the region defined by expression (48), that is, the feasible set of problem (24)–(28) is empty. In actual applications, what one can do in this case is either to increase the budget or to slacken condition (26) by allowing for some overcrowding in the facilities. This can be done by formulating the mathematical programming problem:

$$\text{minimize } \frac{1}{\beta} \sum_{i,j} S_{ij} \log S_{ij} + \sum_{ij} S_{ij} C_{ij} , \tag{49}$$

subject to

$$\sum_j S_{ij} = P_i , \tag{50}$$

$$\sum_i S_{ij} = \rho x_j , \tag{51}$$

$$\sum_j a_j x_j = R , \tag{52}$$

$$x_j \geqslant 0 , \tag{53}$$

where ρ is the maximum allowable density, as measured by the ratio of the number of users to facility size. Problem (49)–(53) can be reduced to the same form as

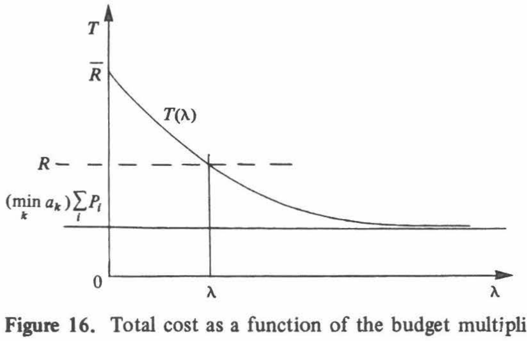

Figure 16. Total cost as a function of the budget multiplier.

problem (36)–(38). If one eliminates the x_j variables by means of constraint (51) and defines $a_j^* = a_j/\rho$, one gets the problem

$$\text{minimize } \frac{1}{\beta}\sum_{i,j} S_{ij}\log S_{ij} + \sum_{i,j} S_{ij}C_{ij} , \tag{54}$$

subject to

$$\sum_j S_{ij} = P_i , \tag{55}$$

$$\sum_j a_j^* \sum_i S_{ij} = R , \tag{56}$$

which is the same as problem (36)–(39), if a_j is replaced by a_j^*.

A common question to be asked in applications is how large the density ρ must be in order to meet the budget constraint. The minimum value of ρ can be found by imposing the condition

$$R \geqslant \left(\sum_i P_i\right) \min_j \frac{a_j}{\rho} , \tag{57}$$

which ensures a nonempty feasible set. Solving for ρ gives

$$\rho \geqslant \min_j \frac{a_j \sum_i P_i}{R} . \tag{58}$$

The right-hand term of expression (58) is the ratio between the cost of establishing a facility large enough to have no congestion and the actual resources one can use to establish the facility. When this ratio is greater than one even in the cheapest location, it gives a measure of the least trade-off that is necessary between congestion and resources.

3.2.2 *Elastic demand case.* It will now be shown how a special form of demand sensitiveness to accessibility can be introduced, which is consistent with a general Neuburger consumer surplus function.

The proposed general form is:

$$\text{minimize } \frac{1}{\beta}\left[\sum_i (P_i - G_i)\log(P_i - G_i) + \sum_{ij} S_{ij}\log S_{ij} + \sum_i (P_i - G_i)C_{i0} + \sum_{ij} S_{ij}C_{ij}\right] , \tag{59}$$

subject to

$$\sum_j S_{ij} = G_i , \tag{60}$$

$$\sum_i a_j x_j \leqslant R , \tag{61}$$

$$x_j \geqslant 0 , \tag{62}$$

(the implicit constraints $S_{ij} \geqslant 0$ and $G_i \leqslant P_i$ are automatically accounted for by the form of the objective function) where the definitions of the variables are the same as for the fixed-demand model (24)–(28), except for the new ones:

G_i is the total actual demand in i (not fixed, but to be found); and

C_{i0} is a parameter which can be interpreted as the 'cost' associated with having no destination, that is, not being a user of any facility.

It must be stressed that, unlike in the fixed demand model, here the total actual demand which will be generated in each demand location is not fixed and known in advance, but has to be determined as a function of the offered service itself. However, it is still assumed that an upper bound P_i is known for the demand in location i, and actual demand G_i will be generally less than P_i. This upper bound can be interpreted as the potential demand in i, that is, the demand which would be generated with infinite accessibility to the service. The special form of the 'entropy' term in the objective function (59) has been adopted to embed both the flow (or allocation) variables S_{ij} and the generated demands G_i.

The elastic demand assumption is probably not the best suited one for those services which satisfy real needs (for instance, primary schools or emergency medical care). It is certainly sound in many other services, however, such as high schools, libraries, recreational services, and also some health care facilities such as hospitals.

In analogy with the fixed demand model, it can be shown how problem (59)-(62) is a general formulation which includes other models as special cases. If, for example, only all-or-nothing values for the x_j variables and nearest-facility allocations are allowed, one obtains a special form of the price-sensitive demand location models discussed in Wagner and Falkson (1975) and in Erlenkotter (1977). It is shown in these references how some changes of variables bring this problem back to the standard form of the plant-location model, so that the usual algorithms can be used. Let us now go back to the more general model (59)-(62), and make a very simple, but fundamental observation.

From the definition of the variables $S_{i0} = P_i - G_i$, the assumption $a_0 = 0$, and with summations over j starting from 0, it is easily seen that problem (59)-(62) is equivalent to the following problem:

$$\text{minimize } \frac{1}{\beta}\sum_{i,j} S_{ij} \log S_{ij} + \sum_{ij} S_{ij} C_{ij} ,$$

subject to

$$\sum_j S_{ij} = P_i , \qquad \sum_i S_{ij} \leqslant x_j , \qquad \sum_i a_j x_j \leqslant R , \qquad x_j \geqslant 0 .$$

A comparison with problem (24)-(28) shows that the two problems are completely identical. What has been done is to introduce a dummy location, labeled by $j = 0$, where a dummy facility will be located at no cost. The dummy facility will be made big enough to serve all the potential demand which is not served by the real facilities.

The above observation means that all the theoretical and computational problems related to model (59)-(62) can be solved by the same approaches developed for the fixed-demand model, except for only slight changes. However, the main results will be restated in new terms for convenience. In what follows, summations over j will be again with $j \neq 0$.

The following definitions are made:

$w_j = \exp(-\lambda a_j)$ the attraction weight in location j (where λ is the Lagrange multiplier for the budget constraint);

$f_{i0} = \exp(-\beta C_{i0})$ where C_{i0} are given constants;

$f_{ij} = \exp(-\beta C_{ij})$ the exponential decay factor for a travel cost C_{ij}; and

$\phi_i = \sum_j f_{ij} w_j$ the Hansen (1959) measure of accessibility to the service from the demand point i.

Then one has:

$$G_i = P_i \frac{\phi_i}{\phi_i + f_{i0}} , \tag{63}$$

$$S_{ij} = G_i \frac{f_{ij} w_j}{\phi_i} , \tag{64}$$

$$x_j = \sum_i S_{ij} = \frac{w_j}{\phi_i} \sum_i G_i f_{ij} . \tag{65}$$

From function (63) it can be seen how the elastic demand assumption works. The total actual demand generated in i, G_i, is a function of the accessibility to the service from i; it is zero for zero accessibility (which means no available facilities) and tends to the total potential demand P_i for infinite accessibility. The shape of function (63) is shown in figure 17, which should be compared with figure 18 (which is the same as figure 7 in part 1 of the paper).

We conclude this section by noting that, unlike in the fixed demand case, here the feasible region is nonempty for any $R \geqslant 0$, since the unsatisfied potential demand is automatically pushed out of the system by the elastic-demand mechanism implied by function (63).

3.2.3 *Models without capacity constraints.* It has already been stressed that the models discussed in 3.2.1 and 3.2.2 are doubly constrained, because they have constraints both on the total demand coming from each location and on the total number of users that can be served at each location. Equality between attracted demand and available capacity in each location is always required. If this assumption is dropped, another family of location models is obtained. These models, first introduced for the location of retail trade (Harris, 1964; Lakshmanan and Hansen, 1965; Huff, 1964; 1966) assume that the users' behavior is described by a spatial-interaction model with no constraints on the destinations, where the attractiveness of the facilities is measured by a nondecreasing function of their size (usually a power function). Although these models were mainly used for retail trade location, it is argued that they might be useful for some kinds of public services too.

Recently, mathematical programming formulations, related to the ones already discussed in sections 3.2.1 and 3.2.2, have been developed for these models. They are briefly discussed here.

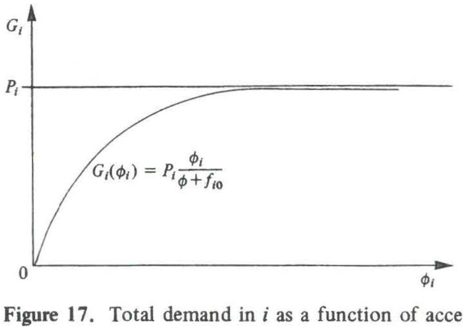

Figure 17. Total demand in i as a function of accessibility.

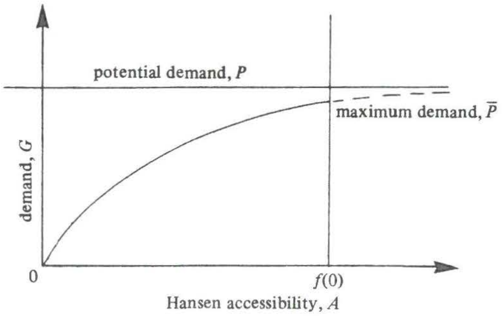

Figure 18. An accessibility-increasing demand curve.

The usual general form for these models is (Coelho and Wilson, 1976):

$$\underset{\mathbf{S},\,x}{\text{maximize}} \sum_{i,j} S_{ij}\left(\frac{\alpha}{\beta}\log x_j - C_{ij}\right) - \frac{1}{\beta}\sum_{i,j} S_{ij}\log S_{ij} \,, \tag{66}$$

subject to

$$\sum_j S_{ij} = P_i \,, \tag{67}$$

$$\sum_j a_j x_j = R \,, \tag{68}$$

$$x_j \geqslant 0 \,. \tag{69}$$

Another form which has been shown to be equivalent to model (66)–(69) and which contains the x_j variables only, is given by Leonardi (1973; 1978):

$$\underset{x}{\text{maximize}} \sum_i P_i \log \sum_j x_j^{\alpha} \exp(-\beta C_{ij}) \,, \tag{70}$$

subject to

$$\sum_j a_j x_j = R \,, \tag{71}$$

$$x_j \geqslant 0 \,. \tag{72}$$

In both cases, the general solution must satisfy the conditions (see the quoted references for the proof):

$$\frac{\sum_i S_{ij}}{\sum_i P_i} = \frac{x_j a_j}{R} \,, \tag{73}$$

where S_{ij} is the flow of users from i to j, and is given by the following production-constrained spatial-interaction model:

$$S_{ij} = P_i \frac{x_j^{\alpha}\exp(-\beta C_{ij})}{\sum_j x_j^{\alpha}\exp(-\beta C_{ij})} \,. \tag{74}$$

Condition (73) states that the budget must be allocated to locations in proportion to the total demand they attract. This balancing principle, more or less explicitly states, has been widely used over a long period of time. A recent review of the concept, its applications, and extensions can be found in Harris and Wilson (1978). Equations (73) can be easily solved by a first-order iteration technique over the x_j.

It is interesting to compare the main features of problem (66)–(69) with the ones of problem (23)–(27). In the more meaningful case the equality holds in expression (35), so that for problem (23)–(27) one has that $\sum_i S_{ij} = x_j$. But this condition will never be like equations (73), unless $R = \sum_i P_i$ and all a_j are equal, that is, there are no differences among the location costs.

Another important difference is found in the sensitivity of the relative size of facilities to changes in the budget. Indeed, model (66)–(69) is easily seen not to be sensitive at all, since it keeps the same relative distribution of sizes and locations for all the possible budget values, and no concentration occurs because of the scarcity of resources.

3.2.4 *Optimal location with a partially existing stock.* In the models discussed in the preceding sections no distinction is made between different types of costs. A general overall unit cost is assumed for the facilities, which will possibly include both building and running costs. Also, it is assumed that both costs are always met for every facility; this is the same as assuming that new facilities are always built and no account is taken of existing ones, if any.

In a more realistic setting, however, locational decisions are made based on an already built environment, where an existing stock of facilities is usually available. In the more general case, decisions have to be made on expanding the existing capacity by building new facilities, by using part of the existing capacity, and by closing down some of the existing facilities. Each one of these actions implies different kinds of costs, and in what follows an attempt will be made to take this into account.

It should also be stressed beforehand that location problems with partially existing facilities lead naturally to dynamic formulations. The models discussed in this section can thus be looked upon as a first step towards dynamic generalizations. The third kind of costs (demolition costs) will not be introduced, since demolition is only justified in a fully dynamic formulation of the problem. However, most of the results can be easily generalized in order to account for this problem if the option of demolishing part of the unused capacity is added.

Two main models are proposed: the first with charges added that are a simple linear function of building costs, and the second with fixed charges in building costs added.

1. For a linear function of building costs, and with a general objective function as in function (23), one has the problem

$$\text{minimize } \frac{1}{\beta}\sum_{i,j} \log S_{ij} + \sum_{ij} S_{ij} C_{ij} \,, \tag{75}$$

subject to

$$\sum_{j} S_{ij} = P_i \,, \tag{76}$$

$$\sum_{i} S_{ij} \leqslant y_j + z_j \,, \tag{77}$$

$$\sum_{j} a_j(y_j + z_j) + \sum_{j} b_j z_j + \sum_{j} d_j(v_j - y_j) \leqslant R \,, \tag{78}$$

$$0 \leqslant y_j < v_j \,, \tag{79}$$

$$0 \leqslant z_j \,, \tag{80}$$

where

y_j is the existing capacity in use at location j,
z_j is the newly built capacity in location j,
v_j is the total existing capacity in location j,
b_j is the unit building cost in location j, and
d_j is the unit cost for holding unused capacity in location j.

The following properties of the general solution to problem (75)–(80) are easily proved by means of the Kuhn–Tucker optimality conditions.

(a) The solution always has the form

$$z_j = 0 \,, \qquad \text{if} \quad 0 \leqslant y_j < v_j \,,$$

$$z_j \geqslant 0 \,, \qquad \text{if} \quad y_j = v_j \,.$$

That is, it is needless to expand the capacity of the service where the existing one is not completely used.

(b) If for any location j one has $d_j > a_j$ then y_j must equal v_j. That is, wherever holding unused capacity is more costly than using it, the total existing capacity is fully used.

(c) When constraints (77) and (78) are binding, one has for the flow variables

$$S_{ij} = P_i \frac{w_j f_{ij}}{\sum\limits_j w_j f_{ij}} , \tag{81}$$

where

$$f_{ij} = \exp(-\beta C_{ij}) , \qquad w_j = \exp(-\lambda u_j) ,$$
$$u_j = \begin{cases} a_j - d_j , & \text{if} \quad 0 \leqslant y_j < v_j , \\ a_j + b_j , & \text{if} \quad y_j = v_j , \end{cases} \tag{82}$$

and λ is the Lagrange multiplier associated with the budget constraint (78).
(d) According to equation (81), the total size of the facility in each location is given by

$$x_j = \sum_i S_{ij} = \sum_i P_i \frac{w_j f_{ij}}{\sum\limits_j w_j f_{ij}} , \tag{83}$$

and one also has

$$y_j = \begin{cases} x_j , & \text{if} \quad x_j \leqslant v_j , \\ v_j , & \text{if} \quad x_j > v_j , \end{cases} \tag{84}$$

$$z_j = \begin{cases} 0 , & \text{if} \quad x_j \leqslant v_j , \\ x_j - v_j , & \text{if} \quad x_j > v_j , \end{cases} \tag{85}$$

for the old and new stock, respectively.

As for the simpler model (23)–(27), the solution for nonbinding constraints is found by putting $\lambda = 0$. The minimum value for R under which the feasible set is empty can be found by concentrating as much demand as possible in the locations with the lowest a_j values (using the existing capacity as much as possible), and the remaining demand in the locations with the lowest building costs.

Formulas (81)–(85) can be usefully compared with formulas (39)–(44) in section 3.2.1. The general form of the solution is basically the same, the only difference being in the cost terms (82). Unfortunately, one does not know a priori the locations where constraint (79) is binding on the right side, so that the sensitivity to changes in the total budget is less easily analyzed.

2. For building cost with fixed charges, and with a general objective function as in function (23), one has the problem

$$\text{minimize } \frac{1}{\beta} \sum_{ij} S_{ij} \log S_{ij} + \sum_{ij} S_{ij} C_{ij} , \tag{86}$$

subject to

$$\sum_j S_{ij} = P_i , \tag{87}$$

$$\sum_i S_{ij} \leqslant y_j + z_j , \tag{88}$$

$$\sum_j a_j(y_j + z_j) + \sum_j b_j z_j + \sum_j d_j(v_j - y_j) + \sum_j h_j \delta_j \leqslant R , \tag{89}$$

$$0 \leqslant y_j \leqslant v_j , \tag{90}$$

$$0 \leqslant z_j \leqslant M\delta_j , \tag{91}$$

$$\delta_j \in \{0, 1\} , \tag{92}$$

where

h_j is the fixed charge to build new facilities in location j,

δ_j is a Boolean variable, which is set to 1 when capacity in j is expanded, and 0 otherwise, and

M is a very large number.

Model (86)-(92) looks very much like model (75)-(80), except for the new costs and the δ_j variables. Provided the values for the δ_j are known, the model has the same form and can be treated in the same way. Exact and/or heuristic methods to find the optimal δ_j values are therefore needed. This can be done with methods related to the ones developed for the classical plant-location problem. (The special structure of problem (86)-(92) has still to be exploited, however.) This is an area in which new research has still to be carried out.

An elastic-demand version of models (75)-(80) and (86)-(92) can be developed with the same approach used in model (66)-(69). If the objective function (66) or (70) replaces objective function (75) and constraint (60), it is easily shown that for the optimal solution one must have

$$\sum_i S_{ij} = \begin{cases} \lambda(a_j - d_j)y_j \,, & \text{if} \quad y_j < v_j \,, \quad z_j = 0 \,, \\ \lambda(a_j + b_j)(v_j + z_j) \,, & \text{if} \quad y_j = v_j \,, \quad z_j > 0 \,, \end{cases} \tag{93}$$

where λ is a Lagrange multiplier.

3.3 *Nonlinear budget constraints*

3.3.1 *The basic model.* If nonlinear cost functions replace the $a_j x_j$ terms in constraints (27), model (24)-(28) assumes the following form:

$$\text{minimize } \frac{1}{\beta}\sum_{i,j} S_{ij} \log S_{ij} + \sum_{ij} S_{ij} C_{ij} \,, \tag{94}$$

subject to

$$\sum_j S_{ij} = P_i \,, \tag{95}$$

$$\sum_i S_{ij} \leqslant x_j \,, \tag{96}$$

$$\sum_j f_j(x_j) \leqslant R \,, \tag{97}$$

where $f_j(x_j)$ is the cost of a facility of size x_j in location j, with $f_j(0) = 0$. The functions $f_i(\cdot)$ are assumed to be as smooth as needed and to be concave nondecreasing. This means that scale economies are introduced in the model.

By means of the Lagrange-multiplier method, it is found that the general solution to problem (94)-(97) when constraints (96) and (97) are binding satisfies the following equations:

$$S_{ij} = P_i \frac{h_{ij} w_j}{\phi_i} \tag{98}$$

for the flow variables, where by definition

$$h_{ij} = \exp(-\beta C_{ij}) \,, \qquad w_j = \exp[-\lambda f_j'(x_j)] \,, \qquad \phi_i = \sum_j w_j h_{ij} \,,$$

$f_j'(x_j)$ is the first derivative of $f_j(x_j)$, and λ is the Lagrange multiplier associated with constraint (97).

Equations (97) can be compared with expression (39) for the analogy; however, equations (98) are not a closed-form solution, since the right-hand side is a function of the x_j variables. For these one can write the equations

$$x_j = F_j(x) \tag{99}$$

where

$$F_j(x) = \sum_i S_{ij} = \sum_i P_i \frac{h_{ij} w_j}{\phi_i} \,, \tag{100}$$

and x is the $[x_j]$ vector. The right-hand side of (97) can be rearranged in a more meaningful way

$$S_{ij} = \frac{P_i}{1 + B_{ij} \exp[\lambda f_j'(x_j)]} \,, \tag{101}$$

where by definition

$$B_{ij} = \sum_{k \neq j} \frac{h_{ik}}{h_{ij}} \exp[-\lambda f_k'(x_k)] \,, \tag{102}$$

a term which does not depend on x_j. Formula (101) is a logistic in $f_j'(\cdot)$, so that one has for the derivatives

$$\frac{\partial S_{ij}}{\partial x_j} = \lambda g_j S_{ij}(P_i - S_{ij}) \,, \tag{103}$$

where the $g_j(x_j)$, $g_j(x_j) = -f_j''(x_j)$, are the second derivatives of the cost curves multiplied by -1; $g(\cdot)$ is always nonnegative because of the concavity assumption. The derivatives of the functions $F_j(\cdot)$ defined by expression (100) are thus given by

$$\frac{\partial F_j}{\partial x_j} = \lambda g_j \sum_i S_{ij}(P_i - S_{ij}) \,. \tag{104}$$

A further requirement has to be met by the functions $f_j(\cdot)$. Of course, one must have $S_{ij} = 0$, if $x_j = 0$. Formula (101) shows that this happens only if

$$f_j'(0) = \infty \,, \tag{105}$$

that is, the cost curves have to be infinitely steep at the origin. This is a kind of smooth version of fixed-charge costs, which introduce scale economies by means of an infinitely steep initial jump. When condition (105) is not met, one can never have a solution with $x_j = 0$ for some j; the linear-cost models are a particular example of this general fact.

Research has to be carried out to find convenient numerical solution methods for formula (99), possibly with special assumptions on the functions $f_j(\cdot)$. The simplest assumption that can be made is that of constant elasticities, so that the $f_j(\cdot)$ will be power functions. They can be written in the following form:

$$f_j(x) = \frac{a_j x^{1-\alpha_j}}{1-\alpha_j} \,, \qquad 0 < \alpha_j < 1 \,,$$

so that

$$f_j'(x) = a_j x^{-\alpha_j} \,. \tag{106}$$

The following set of equations for the variables x_j and the multiplier λ should be solved:

$$x_j = \sum_i P_i \frac{h_{ij} \exp(-\lambda a_j / x_j^{\alpha_j})}{\sum_j h_{ij} \exp(-\lambda a_j / x_j^{\alpha_j})} \,, \tag{107}$$

$$\sum_j \frac{a_j}{1-\alpha_j} \left[\sum_i P_i \frac{h_{ij} \exp(-\lambda a_j / x_j^{\alpha_j})}{\sum_j h_{ij} \exp(-\lambda a_j / x_j^{\alpha_j})} \right]^{1-\alpha_j} = R \,. \tag{108}$$

This approach can be also used to build a nonlinear-cost version of models (66)–(69) and (70)–(72), with size-dependent utility functions. One easily obtains for the optimal point the conditions

$$\frac{\sum\limits_i S_{ij}}{\sum\limits_i P_i} = \frac{x_j f_j'}{\sum\limits_j x_j f_j'} \, ,$$

which can be usefully compared with conditions (76). Elastic demand can be introduced as in section 3.2.2 and needs no new formal tools.

However, it must be recalled that the assumption of concavity for the cost functions destroys the uniqueness property of solutions to problem (94)–(97), and multiple nonoptimal solutions may be found. The usefulness of equations (107) and (108) for finding numerical solutions is therefore limited, and must still be explored.

3.3.2 *Nonlinear cost functions with partially existing stock.* The proposed model is analogous to the ones introduced in section 3.2.4:

$$\text{minimize} \ \frac{1}{\beta}\sum_j S_{ij}\log S_{ij} + \sum_j S_{ij}C_{ij} \, , \tag{109}$$

subject to

$$\sum_j S_{ij} = P_i \, , \tag{110}$$

$$\sum_i S_{ij} \leqslant y_j + z_j \, , \tag{111}$$

$$\sum_j f_j(y_j+z_j) + \sum_j g_j(z_j) + \sum_j h_j(v_j - y_j) \leqslant R \, , \tag{112}$$

$$0 \leqslant y_j \leqslant v_j \, , \tag{113}$$

$$0 \leqslant z_j \, , \tag{114}$$

where the variables are defined as in section 3.2.4, and:
$f_j(\cdot)$ is the running-cost function for location j,
$g_j(\cdot)$ is the building-cost function for location j,
$h_j(\cdot)$ is the unused-capacity-cost function for location j.

The cost functions are assumed concave nondecreasing, so that scale economies are introduced.

The following properties of the general solution to problem (109)–(114) are easily proved.

(1) The solution always has the form $z_j = 0$, if $0 \leqslant y_j < v_j$, and $z_j > 0$, if $y_j = v_j$, as in the linear-cost case.

(2) If the inequality $h_j(v_j) \geqslant f_j(v_j)$ holds, that is, if the cost for holding the empty stock v_j is greater than the cost for using it, then one always has $y_j = v_j$, so that location j will always have a facility size at least as great as the existing stock.

(3) If the inequalities $h_j(v_j) < f_j(v_j)$ and $h_j'(0) > f_j'(v_j)$ hold, then the equation $f_j(y_j) + h_j(v_j - y_j) = f_j(v_j)$ has a root in $(0, v_j)$, where $0 < \overline{y}_j < v_j$, and in location j we will have either $0 \leqslant y_j \leqslant \overline{y}_j$ or $y_j = v_j$.

(4) If the inequalities $h_j(v_j) < f_j(v_j)$ and $h_j'(0) < f_j'(v_j)$ hold, in location j one will have $0 \leqslant y_j \leqslant v_j$.

(5) When constraints (111) and (112) are binding, one has for the flow variables

$$S_{ij} = P_i \frac{\psi_{ij} w_j}{\phi_i}$$

where

$$\psi_{ij} = \exp(-\beta C_{ij}) , \qquad \phi_i = \sum_j \psi_{ij} w_j , \qquad w_j = \exp[-\lambda\mu_j(x_j)] , \tag{115}$$

$$\mu_j(x_j) = \begin{cases} f_j'(x_j) - h_j'(v_j - x_j) , & \text{if } \ 0 \leqslant x_j \leqslant v_j , \\ f_j'(x_j) + g_j'(x_j - v_j) , & \text{if } \ v_j < x_j , \end{cases} \tag{116}$$

with λ the Lagrange multiplier associated with constraint (112), and x_j the total size of the facility in location j.

If $f_j(x_j)$ is replaced by $\mu_j(x_j)$ as defined in expression (116), equations (99)–(104) still hold. The necessary condition to have a zero-size facility for some location j is still

$$f_j'(0) = \infty , \tag{117}$$

and it is interesting to notice that this depends on running costs only.

Specific instances of the above model can be obtained by introducing explicit forms for the cost functions, as in equations (106)–(108).

3.4 *Towards a more general family of location models*

3.4.1 *The general model.* The models discussed so far have a rather specialized objective function, rooted in the Neuburger consumer surplus maximizing and spatial-interaction theory. It has been shown that this form is general enough to include linear-integer plant-location models as special cases. It has also been shown that the introduction of appropriate new variables allows for a particular form of elastic demand.

A new form is now introduced for the objective function, in which the entropy term is replaced by a more general users' benefit function. It is assumed that the total benefit is made up of three terms as follows:

$F_{ij}(S_{ij})$ a term depending on the flows between i and j only,
$E_i(G_i)$ a term depending on the total demand coming from i ($G_i = \sum_j S_{ij}$), and
$H_j(x_j, D_j)$ a term depending on the size of the facility x_j and on the total demand it attracts ($D_j = \sum_i S_{ij}$).

Adding up all the terms over i and j and subtracting the usual travel cost gives a total net benefit function of the following form:

$$B = \sum_{ij} F_{ij}(S_{ij}) + \sum_i E_i(G_i) + \sum_j H_j(x_j, D_j) - \beta \sum_{ij} S_{ij} C_{ij} . \tag{118}$$

If the objective function (118) is introduced into a general model with partially existing stock and nonlinear cost functions, the following mathematical programming problem is obtained

$$\text{maximize} \sum_{i,j} F_{ij}(S_{ij}) + \sum_i E_i(G_i) + \sum_j H_j(y_j + z_j, D_j) - \beta \sum_{ij} S_{ij} C_{ij} , \tag{119}$$

subject to

$$\sum_j S_{ij} = G_i , \tag{120}$$

$$\sum_i S_{ij} = D_j , \tag{121}$$

$$D_j \leqslant (y_j + z_j) , \tag{122}$$

$$\sum_j f_j(y_j + z_j) + \sum_j g_j(z_j) + \sum_j h_j(v_j - y_j) \leqslant R , \tag{123}$$

$$0 \leqslant y_j , \tag{124}$$

$$y_i \leqslant v_j , \tag{125}$$

$$0 \leqslant z_j , \tag{126}$$

where the usual definitions and assumptions hold. [Constraints (120) and (121) are actually redundant, since the variables G_i and D_j can be eliminated.]

It can be shown that the general solution to problem (119)-(126) has the same properties as properties (1)-(4) listed in section 3.3.2.

3.4.2 *An example.* Specific instances of model (119)-(126) can be obtained in various ways. Here an attempt will be made to show what can be done, with no claim of generality. What will be developed is just one of the ways of looking at model (119)-(126), which is considered to be useful for some applications.

Suppose one has:

$$-F_{ij}(S_{ij}) = \frac{1}{\beta}S_{ij}\log S_{ij} ,$$

the usual entropy term;

$$-E_i(G_i) = \frac{1}{\beta}(P_i - G_i)\log(P_i - G_i) + a_i ,$$

a term similar to the one introduced in the elastic demand model (59)-(62), it can be interpreted as a cost associated with unsatisfied potential demand; and

$$-H_j(D_j) = \frac{\alpha}{\beta}(x_j - D_j)\log(x_j - D_j) + b_j ,$$

a cost term associated with the deviation between usable capacity and actual demand. Then substitution into the objective function (119) yields the mathematical programming problem

$$\text{minimize } \frac{1}{\beta}\sum_{ij} S_{ij}\log S_{ij} + \frac{1}{\beta}\sum_i (P_i - G_i)\log(P_i - G_i) + \frac{\alpha}{\beta}\sum_j (x_j - D_j)\log(x_j - D_j)$$

$$+ \sum_{ij} S_{ij}C_{ij} + \sum_i (P_i - G_i)a_i + \sum_j (x_j - D_j)b_j , \tag{127}$$

subject to

$$D_j \leqslant x_j , \tag{128}$$

$$\sum_j f_j(x_j) + \sum_j g_j(z_j) + \sum_j h_j(v_j - y_j) \leqslant R , \tag{129}$$

$$0 \leqslant y_j \leqslant v_j , \tag{130}$$

$$0 \leqslant z_j , \tag{131}$$

where

$$x_j = y_j + z_j , \qquad G_i = \sum_j S_{ij} , \qquad D_j = \sum_i S_{ij} .$$

The constraint (128) is actually redundant if $\alpha > 0$, but it will be useful to keep it in order to obtain the correct form for the limiting cases. Assuming that constraint (129) is active and proceeding as usual, one obtains the following results for the general solution:

$$G_i = P_i\frac{\phi_i}{v_i + \phi_i} , \tag{132}$$

$$S_{ij} = G_i\frac{w_j\exp(-\beta C_{ij})}{\phi_i} , \tag{133}$$

where

$$\nu_i = \exp(-\beta a_i) \, ,$$

$$w_j = \exp(\alpha\beta b_j)(x_j - D_j)^{\alpha} \, , \tag{134}$$

$$\phi_i = \sum_j w_j \exp(-\beta C_{ij}) \, . \tag{135}$$

For an optimal x one further has

$$w_j = \exp-[\alpha + \lambda\mu_j(x_j)] \propto \exp-\lambda\mu_j(x_j) \tag{136}$$

where $\mu_j(\cdot)$ are the functions defined by expression (116). A comparison with some earlier results is useful. Equations (132) and (133) look like equations (63) and (64), implying that an elastic demand behavior is embedded in the model. Equation (136) is basically the same as equation (115). This means that the solution to problem (127)–(131) is indistinguishable from the solution to problem (109)–(114), as far as the sizes and locations of facilities are concerned. However, a different behavior is implied for the users. From equation (134) it is seen that the attraction weight w_j is proportional to a power of the difference between capacity and attracted demand, and by changing the values of the parameters α and b_j one obtains a different sensitivity to the deterrence effect of congestion. As a special case, if it can be assumed that D_j is always small, compared to x_j, and all b_j are equal to b, the attractiveness weights are given approximately by $w_j \approx kx_j^{\alpha}$, with $k = \exp(\alpha\beta b)$. This assumption implies that the system never produces congestion, so that one obtains models related to the ones discussed in section 3.2.3. If, on the other hand, $\alpha = 0$, the terms in x_j and D_j disappear from the objective function (127), and constraint (128) usually becomes active. The problem thus reduces to an elastic-demand version of model (109)–(114).

In short, it is seen that model (127)–(131) yields all the previously discussed models as special cases, including models with elastic demand and models with size-dependent utility functions. Furthermore, when its parameters do not assume limiting values, its general solution is fairly flexible in fitting users' behavior, although still consistent with the Neuburger consumer surplus formalism.

4 Concluding comments and issues for further research

4.1 *General comments*

A general framework for future research has been outlined in section 2 (last issue, pages 1002–1025) in broad nontechnical terms. The purpose of this section is to add a few technical notes on problems of a less general nature, but of some importance to applications.

Single period (or static), single-level optimal location problems are usually supposed to be rather well-known and solvable. However, it can be seen in the preceding sections how poorly solved or unsolved problems can still be found. This is even more true if the term 'problem' is given a wider meaning than 'optimization problem' and the term 'solution' a wider meaning than 'optimal solution' to a possible mathematical program. Indeed problems of content, form, and algorithms must be solved.

Problems of content include such topics as physical, behavioral, and economic assumptions and interpretations; and consistency between the models and the aggregation levels they are used for. *Problems of form* are concerned with the mathematical tools we use to formulate the problems. They range from continuous (rather than discrete) to nonlinear (rather than linear) to existence and uniqueness analyses. *Problems of algorithms* are concerned with the best ways to solve, or nearly solve, the problems stemming from content and form. Surprisingly enough, it seems as though people working on location problems actually tend to group themselves according to the three categories above, with little exchange of information among the groups.

Unification, rather than new models or algorithms, is thus the main need, and sections 3.1 to 3.4 in this paper were guided by this idea. However, it must be said that unifying features have been sought mainly in the form of problems, whereas comments on their content have been kept intentionally vague. This happens mainly because the unifying of forms is a relatively easy task, whereas the carrying out of a satisfactory analysis of content is the real challenge, and it needs a substantial amount of time and effort.

This is not to say that the models discussed are useless. Indeed, the belief is that static, single-level optimal-location problems not falling within these models are in some way hard to find. What is needed from now on is just to find where each problem falls, and how and why.

The stress placed on contents does not mean there are no algorithmic problems. Actually, the most sensible general-purpose models, like the ones discussed in section 3, basically need new algorithms. More than this, they need a lot of effort on coordinating both existing and possibly new algorithms.

As far as the purely formal problems are concerned, it is believed that no really new mathematical effort is needed for static, single-level problems. If a change is made from static to dynamic and from single to multilevel (or multifacility) problems, this is no longer true. However, dynamic multilevel problems fall outside the purpose of this paper, and will be the subject of future studies.

The first two problems posed above are now discussed in greater depth, in order to draft a framework for future research. The problem of algorithms is too technical for an introductory paper like this. It will be, however, the subject of some forthcoming work.

4.2 *Content*

An optimal location model is made up of an objective function, some constraints and, of course, some decision variables. Since there is no general agreement on what these components are, a brief discussion is necessary. For this purpose it is easier to take them in reverse order.

4.2.1 *The decision variables.* All location problems have size and location of facilities among the decision variables. Some of them also have commodity or users' flows. All of them may be different depending on what is meant by 'location', 'facility', and 'flows'. Because of this it is believed that research effort is needed for the following:

1. *To identify the correct definition of locations for each problem.* This issue is closely related to the aggregation level problem, and particular care must be placed on distinguishing the problems with pointwise locations from the ones where locations are zones, or subregions of a region.
2. *To identify the correct definition of facility for each problem.* This problem is related to the one above, so that the 'size' of a 'facility' in a given location has different meanings depending on what the location is. It is usually a plant, a building, or a single piece of equipment for pointwise locations, whereas it is an aggregate usually of many plants, or of buildings, or of many pieces of equipment in zone-like locations.
3. *To identify the correct definition of flows for each problem.* As stressed many times in sections 2 and 3.1 to 3.4, there is a *fundamental* difference between problems where displacements of goods and people are charged to the users and the ones where displacements are controlled by the locational decisionmaker. Whereas in the latter case the well known OR plant-location models apply (with possible slight

variations for each problem), in the former case models based on spatial-interaction theory are needed. Regrettably enough, mistakes are often made in choosing the model to be used. Even more regrettable is the tendency of people who like one approach to dislike the other one, regardless of the nature of the problem being solved. Actually, one needs to know which approach is best suited for each problem, and not which is 'good' or 'bad'.

4.2.2 *The constraints.* A general agreement exists on the main set of typical constraints for a location problem. Most of them have an obvious physical meaning and need nothing new. The only one requiring further physical and/or economic insight is the constraint on total budget[2].

This is mainly a problem of defining the costs for a single-period location plan. The two main research issues seem to be:

(a) *The problem of defining running and building costs.* Most location models pay little attention to the difference between cost components which have to be summed and discounted over time (running costs for a facility), and the cost components found once in a while (building a new or expanding an old facility). Obviously enough, this mainly happens because of their static nature. However, since most single-period-location models have to be used as steady-state approximations to dynamic problems, it may well happen that continuously discounted running costs prevail over building costs, as far as locational and/or relocational decision are concerned. Therefore a deeper analysis of running-cost components is needed.

Another issue related to the total cost problem is that of demolition. In many location-relocation problems, where decisions have to be taken on existing stock, it may be meaningful to ask whether demolition is better than keeping unused stock. This problem needs a dynamic setting to be solved; therefore, a deeper analysis of this cost component is needed.

(b) *The problem of scale economies.* Introducing convex nondecreasing cost functions is usually realistic for problems with pointwise locations, where it can be certain that 'a facility' is actually a single facility, or plant, or piece of equipment. It is not so sensible for problems where locations are zones or subregions, since in this case it is not known how the total size allocated to each zone is spread across it, how many single facilities it is made up of, how big each single facility is, and so on. In this case, linear costs and no fixed charges are probably the best compromise. Another issue related to scale economies is the tendency toward concentration that these economies usually induce in the models. This is a controversial point, since there are people thinking that a few big facilities are the best solution, and those thinking that social welfare and equity require evenly distributed facilities across space.

4.2.3 *The objective function.* As discussed earlier, different kinds of objective functions can be found. Two important ones are:

(1) Functions where costs (benefits) charged to the locational decisionmaker are minimized (maximized).

(2) Functions where costs (benefits) charged to the users are minimized (maximized).

Whereas the first kind of function poses no really new problem, the second kind does. Provided one accepts the embedding approach, by which maximizing is used both to optimize (a decisionmaker's task) and to induce the appropriate users' behavior, much can be said about what the appropriate users' behavior is, and how it should be embedded in the model.

[2] In many models there is nothing like a budget constraint, since the total cost is part of the objective function. However, there is no difference between the two formulations, as far as the meaning of the cost functions is concerned.

The models which embed the logit-choice behavior are particularly important examples. How these can be generalized and still remain within the overall logit structure has been briefly presented in section 3.4. It could be useful ti find more general functional forms, including the logit as a special case, but allowing for different behaviors as well.

A more general issue concerning the objective function is the relationship between the way the public facility location problem is posed and the economy of the country where it is posed. Although similar in the mathematical form, problems may have very different contents in this respect. The economic interpretation of benefits, for instance, can be some kind of consumer's surplus as has been done in this paper, thus implying the existence of prices. These are usually nonmonetary prices, expressed in terms of users' perceived utilities (or disutilities). This is implied in the random utility interpretation of models based on the logit-behavior assumption (see Neuburger, 1971; McFadden, 1974a; 1974b; Williams, 1977; and van Lierop and Nijkamp, 1979 among others).

A closely related problem is whether public facilities should charge actual direct money prices to the consumers, that is, should they be profit-making, self-financing, or should they just exploit a given budget to maximize social benefit only. In this sense, having total cost in the objective function rather than in a budget constraint makes a difference, although the Lagrangian functions for the two problems look the same.

A third important issue related to the objective function is the problem of empirically fitting the implied consumers' behavior models on actual data. This problem is shared by all the formulations where some kind of elastic or price sensitive demand is assumed, whether these models embed consumers' behavior by Neuburger's consumer surplus maximizing (thus producing standard spatial-interaction models), whether they have demand functions in terms of actual money prices, or whether they use any other theoretical approach.

4.3 *Form*

As previously stated, there are no really new problems as far as the mathematical form of a static location model is concerned. This is especially true for the classical (nonsensitive-demand) plant-location problems, for which a large amount of theoretical and computational work has been carried out, and reference can be made to very good existing state-of-the-art and bibliographical works on that topic (see ReVelle et al, 1970; and Lea, 1979, for instance).

There are, however, some technical problems to be solved for the more general models discussed in sections 3.3.1, 3.3.2, and 3.4, which incidently, turn out to be the most interesting and useful ones in my opinion. They are listed and briefly discussed below:

(a) *Nonconcavity and uniqueness problems.* Unfortunately, the benefit function (to be maximized) is usually concave[3], and the cost functions are also concave. This results in a nonconcave programming problem, whose solution may not be unique (and usually it is not) and whose Kuhn–Tucker points need not be optimal ones. Some existence and uniqueness work could be useful, both for the general model and for its versions assuming meaningful special structures.

(b) *Dual formulation and analysis.* The development of the duals of the general nonlinear models should be carried out in order to fully explore their properties.

[3] If the problem is formulated in terms of cost minimizing, the users' cost function is usually convex, whereas the cost functions in the budget constraint are concave. This results in a nonconvex programming problem.

This should give not only mathematical insight and computational hints, but also better physical and economic interpretations.

(c) *Exploitation of special structures.* Provided enough flexibility is preserved, some meaningful special structures should be introduced in the general functions, both in the objective function and in the budget constraint. 'Entropy' terms and power functions are an example (which has to be fully analyzed as yet). Some effort should be made towards finding other useful functional forms for users' benefits, or costs (other than the entropy form), as well as other functional forms which can best model the scale-economies' effect. Anyway, the theoretical work on special structures should be carried out in close conjunction with the empirical work of fitting actual data to users' behavior (as discussed in section 4.2).

4.4 *Issues for further research*

The main conclusion which can be drawn from sections 2 and 3 is that some model-building work is needed to account for some important features of real facility location problems.

The question of demand sensitiveness to accessibility and congestion is the new leading *descriptive* issue. The unsatisfactory state of the theory and the relevance in applications ask for the development of new effective approaches to this issue. It is therefore proposed as one of the next steps in research on public facility location.

The problem of finding exact and approximate solution techniques to the newly posed location problems is the leading *applied* issue. Expected results in this direction are outlined below:

(1) Efficient ways of handling nonlinear objective functions (based on the spatial-interaction embedding approach).

(2) Efficient procedures to cope with nonsmooth features arising from real problems (such as fixed charges, bounds on feasible sizes, and indivisibilities);

(3) Computational schemes to analyze sensitivity of locational patterns to benefit/cost trade-off changes and travel cost changes. A comparative analysis of numerical results with some real cases will be carried out, in order to test the procedures and techniques outlined above.

The problem of modeling new dimensions in location problems, like those in multilevel multiactivity and multistage (dynamic) systems, is the main new *theoretical* issue. It will also be of great applied value, since multiactivity dynamic models are the best ones suited for real urban service systems. A promising contribution along these lines should be the development of a set of spatial-interaction based indicators, monitoring the changes in the system and pointing out management policies (including capacity expansion, demolition, and relocation) for the stocks of urban service facilities. Of course, this development is not a short-term one, and it also tends to stretch the boundaries of the Public Facilities Location Task. It is felt therefore, that the development of these approaches will lead to suggestions for new inter-disciplinary research on urban systems, which it is hoped will be useful for possible future tasks within an urban management theme.

References

Coelho J D, Wilson A G, 1976 "The optimum location and size of shopping centers" *Regional Studies* **10** 413-421

Erlenkotter D, 1977 "Facility location with price-sensitive demands: private, public, and quasi-public" *Management Science* 24 378-386

Hansen P, Kaufman L, 1976 "Public facilities location under an investment constraint" *Operational Research '75* Ed. K B Haley (North-Holland, Amsterdam)

Hansen W G, 1959 "How accessibility shapes land use" *Journal of the Americal Institute of Planners* 25 73-76

Harris B, 1964 *A Model of Locational Equilibrium for Retail Trade* mimeograph, Penn-Jersey Transportation Study, Philadelphia, Pa

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

Huff D L, 1964 "Defining and estimating a trading area" *Journal of Marketing* **28** 37-48

Huff D L, 1966 "A programmed solution for approximating an optimum retail location" *Land Economics* **42** 293-303

Lakshmanan T R, Hansen W G, 1965 "A retail market potential model" *Journal of the American Institute of Planners* **31** 134-143

Lea A, 1979 "Welfare theory, public goods, and public facility location" *Geographical Analysis* **11** 218-239

Leonardi G, 1973 "Localizzazione ottimale dei servizi urbani (Optimal location of urban services)" *Ricerca Operativa* **12** 15-43

Leonardi G, 1978 "Optimum facility location by accessibility maximizing" *Environment and Planning A* **10** 1287-1305

Leonardi G, 1980 *On the Formal Equivalence of Some Simple Facility Location Models* WP-80-21, International Institute for Applied Systems Analysis, Laxenburg, Austria, presented in the Workshop on Location and Distribution Management, at the European Institute for Advanced Studies in Management, Brussels, Belgium

McFadden D, 1974a "Conditional logit analysis of qualitative choice behavior" in *Frontiers in Econometrics* Ed. P Zarembka (Academic Press, New York) pp 105-142

McFadden D, 1974b "The measurement of urban travel demand" *Journal of Public Economics* **3** 303-328

Neuburger H L I, 1971 "User benefit in the evaluation of transport and land use plans" *Journal of Transport Economics and Policy* **5** 52-75

ReVelle C S, Marks D, Liebman J C, 1970 "An analysis of private and public sector location models" *Management Science* **16** 692-702

ReVelle C, Rojeski P, 1970 "Central facilities location under an investment constraint" *Geographical Analysis* **2** 343-360

van Lierop W, Nijkamp P, 1979 "A utility framework for interaction models" *Sistemi Urbani* **1** 41-64

Wagner J L, Falkson L M, 1975 "The optimal nodal location of public facilities with price-sensitive demand" *Geographical Analysis* **7** 69-83

Williams H C W L, 1977 "On the formation of travel demand models and economic evaluation measures of user benefit" *Environment and Planning A* **9** 285-344

Wilson A G, 1970 *Entropy in Urban and Regional Modelling* (Pion, London)

PUBLICATIONS IN THE PUBLIC FACILITY LOCATION SERIES

1. Giorgio Leonardi, On the Formal Equivalence of Some Simple Facility Location Models. WP-80-21.

2. Tony J. Van Roy and Donald Erlenkotter, A Dual-Based Procedure for Dynamic Facility Location. WP-80-31.

3. Donald Erlenkotter, On the Choice of Models for Public Facility Location. WP-80-47.

4. Giorgio Leonardi, A Unifying Framework for Public Facility Location Problems. WP-80-79.

5. Giorgio Leonardi, A Multiactivity Location Model with Accessibility- and Congestion-Sensitive Demand. WP-80-124.

6. Yuri Ermoliev and Giorgio Leonardi, Some Proposals for Stochastic Facility Location Models. WP-80-176.

7. Girogio Leonardi and Cristoforo Sergio Bertuglia, Optimal High School Location: First Results for Turin, Italy. WP-81-5.

8. Yuri Ermoliev, Giorgio Leonardi, and Juhani Vira, The Stochastic Quasi-Gradient Method Applied to a Facility Location Problem. WP-81-14.

9. Giogio Leonardi, The Use of Random-utility Theory in Building Location-Allocation Models. WP-81-28.