EQUITY, EFFICIENCY, AND ACCESSIBILITY IN URBAN AND REGIONAL HEALTH CARE SYSTEMS

L. Mayhew
G. Leonardi

July 1981
WP-81-102

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

THE AUTHORS

Leslie Mayhew is an IIASA research scholar working within the Health Care Task of the Human Settlements and Services Area. He is on secondement from the Operational Research Unit of the Department of Health and Social Security, UK.

Giorgio Leonardi has been at IIASA since October 1979 on leave from the Polytechnical Institute of Turin, Italy. He has previously been a researcher at the Italian Association for Housing Research (AIRE) and a consultant on urban and regional planning for the Regional Authority of Piemonte.

FOREWORD

The principal aim of health care research at IIASA has been to develop a family of submodels of national health care systems to use by health service planners. The modeling work is proceeding along the lines proposed in the Institute's current Research Plan. It involves the construction of linked submodels dealing with population, disease prevalence, resource need, resource allocation, and resource supply.

This paper considers four different resource allocation criteria for helping to assess the long-term health resource requirements for different areas of a region. They are based on a spatial interaction model that provides a simple method for selecting between different configurations, when population size and structure and resource availability are changing over time and space. The allocation criteria are based on objectives about which there is broad agreement among planners and other actors in the system. These criteria are concerned with improving the equity or the efficiency of the system, or the accessibility of the population to the supply of health services.

Related publications in the Health Care Systems Task are listed at the end of the paper.

Andrei Rogers
Chairman
Human Settlements
and Services Area

ABSTRACT

This paper explores four different criteria of health care resource allocation at the urban and regional level. The criteria are linked by a common spatial interaction model. This model is based on the hypothesis that the number of hospital patients generated in a residential zone i is proportional to the relative morbidity of i, and to the availability of resources in treatment zone j, but are in inverse proportion to the accessibility costs of getting from i to j. The resource allocation criteria are based on objectives on which there is broad agreement among planners and other actors in a health care system. These objectives are concerned with allocations that conform to notions of equity, efficiency, and two definitions of accessibility. The allocation criteria give mainly aggregate level information, and are designed with the long-term regional planning of health care services in mind. The paper starts by defining the criteria, and describes how they are intended to be employed in a planning context. The allocation rules are then formally derived and linked together mathematically. They are then applied to a region, London, England, which is known to have very complex health care planning problems. As a result of this application, two of the criteria--equity and efficiency--are selected for further analysis. A new model is built and applied that specifically enables the user to trade off one of these criteria against the other.

CONTENTS

EQUITY, EFFICIENCY, AND ACCESSIBILITY IN URBAN AND REGIONAL HEALTH CARE SYSTEMS

1. INTRODUCTION

This paper describes the theory and application of a set of possible methods to assist in the regional planning of health care services. These methods are concerned with finding a set of resource allocations in different parts of a region when the morbidity, demographic structure, and resource availability are changing over time and space. They were designed with applications in the strategic planning of health services in mind, where the decision makers are concerned mainly with the broad directions and outputs of the system over a period of time. The work presented forms part of a wider research effort being carried out both jointly and independently by the Health Care and Public Facility Location Tasks at IIASA (the former also in conjunction with the Operational Research Services of the Department of Health and Social Security, UK). The models that underlie this research are connected by a common spatial interaction methodology (e.g., Wilson 1974), but each is designed to address a slightly different problem either in the health or other public sectors. The level of detail in these models varies according to the intended use and the decision-making level in the system being studied.

In the present case, the outputs of the model forming the basis for the methods described in this study are highly aggregated, but they are typical of the decision variables used at a regional or supra-regional level. Following a discussion of the hypothesis underlying the approach employed and the reasons for this choice, the methods are developed in detail. Each is designed to pick a set of allocations according to one of four different criteria on which there is either broad acceptance by actors in the health care system or considerable precedence in the literature on planning. Particular concern is taken, however, to ensure that the spatial behavior of the patients is correctly embedded in the allocation mechanisms. As a consequence of this concern and of the empirical tests subsequently carried out, two of the criteria are rejected in favor of the remaining two. The two accepted criteria address the problems of systems equity and systems efficiency, respectively--two objectives that are shown to pull the spatial pattern of regional resource allocation in different directions. The other two address the problems of accessibility. To conclude the paper, a composite method with an enhanced range of applications is developed that specifically allows users to develop scenarios trading off the accepted objectives, one against the other.

1.1 Class of Systems

Not all types or sectors of health care systems will provide valid applications for the methods to be described. For example, in highly market-oriented health care systems, services are rationed by mechanisms other than these criteria, and so regional disparities in provision may not receive priority or be even considered a problem. The systems for which this work may be appropriate will probably be drawn from the following types:

- Payment-free or part-payment systems operating comprehensive health insurance schemes where there are few market signals to regulate supply and demand

- Systems with national, regional, or local health care planning machinery and a commitment to the effective territorial planning of health care services

- Systems in which there is a historical tendency to over-allocate resources in some areas and to under-allocate them in others and in which there is a growing desire by statutory authorities to redress these imbalances

- Incipient systems in developing countries, or systems changing from a market approach to a more planned approach in health care delivery in which considerable reorganization may be required

In fact, the applications in this paper are based on data from the United Kingdom, which has operated a nationalized health care system since 1948. The administrative machinery for regional planning, however, has only been in existence since 1973 following reorganization.

1.2 Class of Model

The basic model is formed from the following simple hypothesis. It is that the number of patients generated in an origin zone i (place of residence) and treated in a destination zone j (place of treatment) is in proportion to the morbidity or "patient generating potential" of i and to the resources available in j, but is in inverse proportion to the accessibility costs of getting from i to j. In its current form, the model assumes that there are not enough resources to satisfy demand and that patients are not restricted by their places of residence to use only certain facilities. The first assumption reflects a view (analyzed in more detail below) that whatever is provided tends to get used. The second is to make it clear that only non-emergency services in the acute sector of the health care system are being discussed, and that some freedom of choice as between different facilities is permitted. The type of model that emerges is a gravity model of the attraction constrained form (Wilson 1971).

The model is now stated informally; later it will be derived from theoretical grounds. It is

$$T_{ij} = B_j D_j W_i f(\beta, c_{ij}) \qquad (1)$$

where

$$i = \overline{1,I} \quad , \quad j = \overline{1,J} \quad ,$$

the number of origin and destination zones, respectively, and

T_{ij} = the predicted patient flow from zone i to treatment zone j

D_j = a resource measure defined as the caseload capacity in j for treating patients in a specialty or groups of specialties

W_i = a patient generating factor (pgf), which is an index of the propensity of the population in i to generate patients in the same group of specialties

$f(\beta, c_{ij})$ = a spatial discount function such as $e^{-\beta c_{ij}}$ (as used here) or $d_{ij}^{-\beta}$, which is strictly monotonically declining. Later, this function is abbreviated to f_{ij}

β = a spatial discount parameter (≥ 0) to be determined empirically

c_{ij} = the accessibility costs between i and j

and where

$$B_j = \left[\sum_i W_i f(\beta, c_{ij}) \right]^{-1} \qquad (2)$$

Equation (2) is a constraint that ensures

$$\sum_i T_{ij} = D_j$$

This is the assumption that all resources in j will be used.

Whereas this model ignores the sometimes complex procedures by which patients are referred between different levels and places of treatment in the system, research has shown that it is possible to describe *and* predict accurately the resulting spatial patterns of patient flows between different i and j (Mayhew and Taket 1981), suggesting that the model assumptions are sufficient for its intended purposes. The empirical basis for the model, its range of applications, calibration, and various extensions are given elsewhere (Mayhew and Taket 1980; Mayhew 1980, 1981).

1.3 Mode of Use

In conventional usage, the model predicts the impact on patient flows and hospitalization rates that result from changes in patient generating potential and resource configuration. This permits the evaluation of many alternative allocations, yet it cannot tell the user which is best. For small problems at the local level of decision making, these alternatives will be few, and it is probable that they can be judged for their suitability in only a few computer runs. The strategic level of planning, however, is concerned with the direction of the entire system over a period of time, say 10 to 15 years (DHSS 1976). If a typical planning region contains one or more cities, several towns, over one hundred hospitals, and a service population in excess of ten million, say, the alternative allocations will be too many to evaluate, and the planner will find it useful to direct his search. The methods described here are designed to assist in this search by narrowing down the possibilities to those that in some sense can be judged best and that can be accomplished during the duration of the plan. To do this, however, the model must be directed to pick resource configurations that satisfy a particular objective or set of objectives. The problem is which objectives to choose and how to express them in a way that can be used by the model.

2. THE MAIN OBJECTIVES OF THE HCS

Clearly, a health care system has many objectives, not all of which can be achieved simultaneously. Some objectives, too, will be less important than others, but nevertheless they must be taken into account in some sense (section 2.3). The problem is to understand what the dominant objectives are. It is worth examining the expressed aim of the National Health Service in England and Wales. It is

> ...to ensure that every man and woman and child can rely on getting all the advice and treatment and care they need in matters of personal health...[and] ...that their getting these should not depend on whether they can pay for them (Feldstein 1963:22; quoting from HMSO, 1944).

This seems an uncontroversial statement for the health care systems we have in mind. At least, two serious problems, however, are associated with the ideals expressed in it that are preventing its objectives from being attained. The first is that, as long as patients pay in time, money, discomfort, and other costs for access to facilities, there will always be a negative influence in the volume of per capita health care consumption in different areas no matter which country or what type of health care system is considered. The second is that the assumption in 1944 that all needs could be catered to has proved unrealistic. The budget for health care and the consumption of health care services in general, continues to rise at an alarming rate in the majority of countries, not only in England and Wales. In all countries too, it has proved impossible to measure at a general level the marginal benefits of this increased expenditure, to determine the extent to which genuine needs are being satisfied, or to define an objective set of standards on which to base supply.

2.1 Demand and Availability

Figure 1 illustrates empirically what usually happens in practice when there are uncertainties about outputs, accessibility costs to pay, and excess demands in the system. The discharges and deaths per thousand *catchment* population* (the population mostly dependent on the facilities in an area) are plotted against the hospital bed availabilities in each catchment area in Southeast England in 1977. The diagram demonstrates

(i) the strength of the supply side in the system for determining demand in the areas influenced by the facilities, particularly the way demand seems to rise so that it meets supply**

(ii) the strong dependence of the population on the local availability of facilities

Figure 2 emphasizes point (ii) in another way. It is a histogram showing the relationship between the percentage of patients using facilities in the London area and the distance from the hospital. It is based on a sample of over 2000 patients at 14 hospitals. It shows clearly the marked preference among patients to use local facilities.

2.2 Equity, Efficiency, and Accessibility

Though from the above and other recent evidence, it would appear difficult for a health care system to satisfy all the actual and potential demands for health care, certain criteria stand out as being both sensible and applicable when both budget constraints and uncertain outputs are dominant considera-

---

*A catchment population is defined by C_j where $C_j = \sum_i E_{ij}P_i$, $E_{ij} = T_{ij}/\sum_j T_{ij}$ and P_i is the resident population in i.

**The relationship is not strictly linear since lengths of hospital stay are also an increasing function of bed supply, but this consideration is unimportant in the resource range examined here.

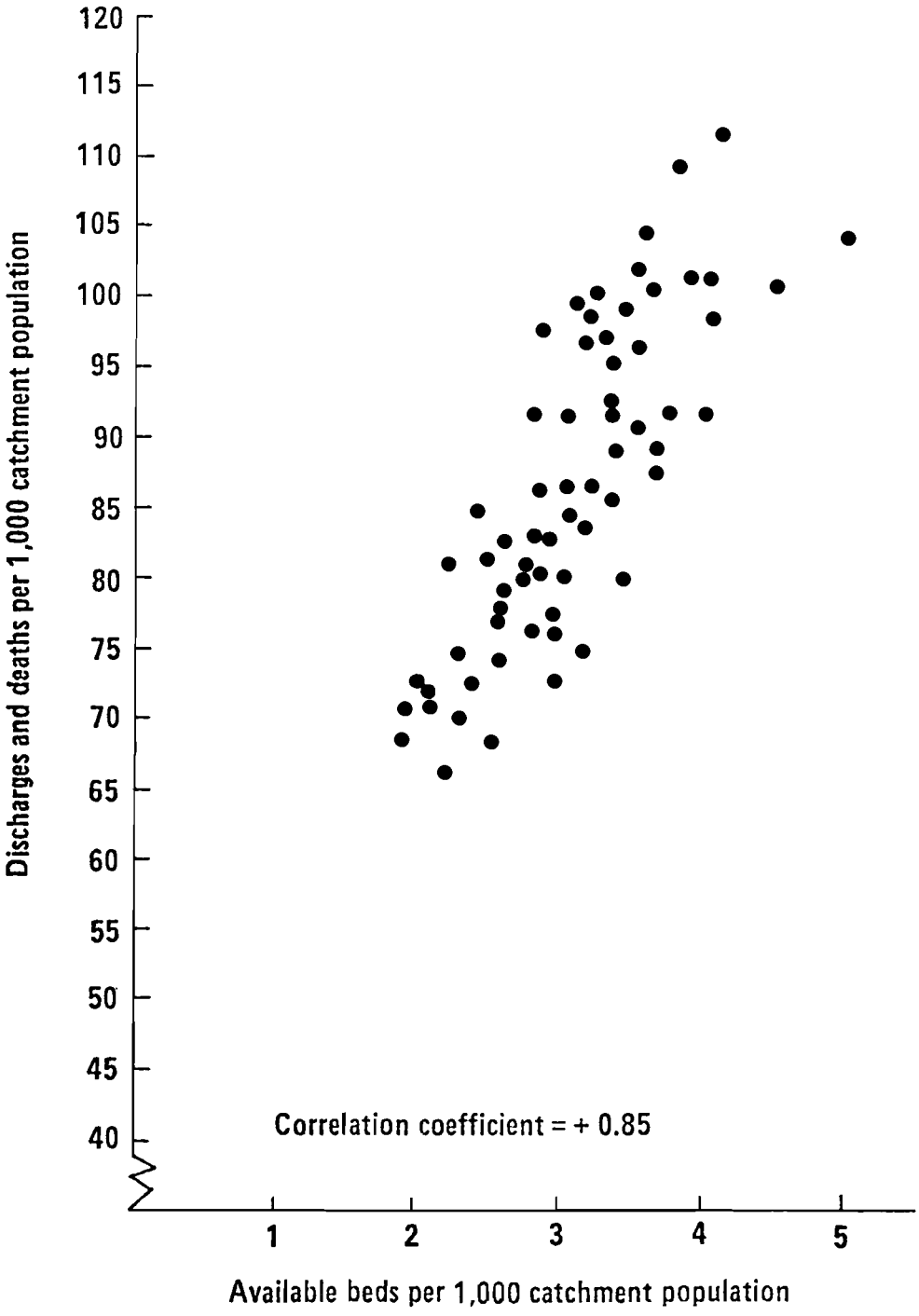

Figure 1. The relationship between hospitalization rates and level of provision for health district catchment populations in Southeast England. (Source: LHPC 1979a:26.)

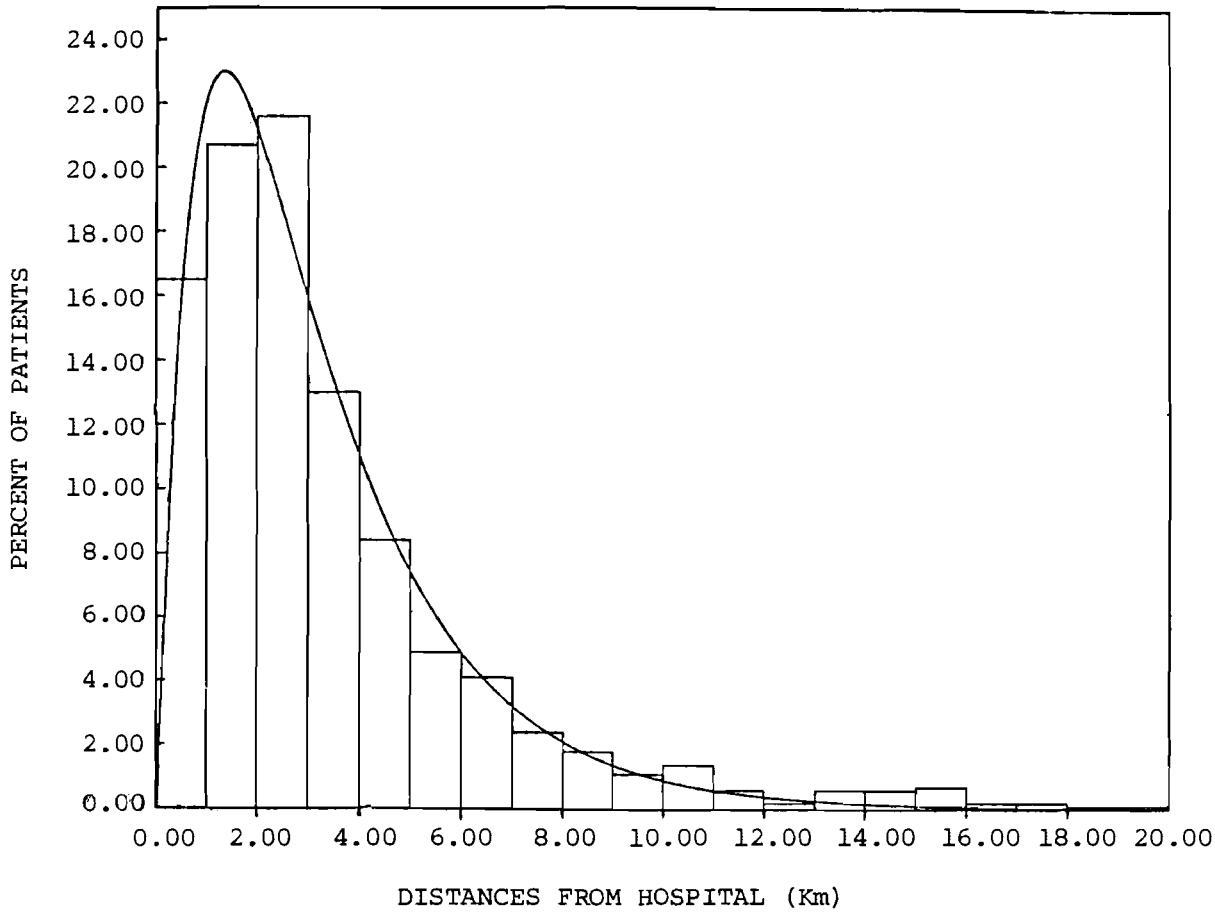

Figure 2. The relationship between the percentage of patients using hospital facilities and the distance from hospitals in London for general medical and surgical specialties. The equation of the fitted curve is $y = 100x^{1.325} \exp(-1.508x^{0.711})$. (Source: Mayhew 1979.)

tions. These criteria are the improvement of the fairness of the system (equity), the increase in benefits to the patients (efficiency), and the equalization of the friction of distance between demand and supply (accessibility).

Equity

The equity criterion is defined as choosing a resource configuration such that the relative needs (not the absolute as above) in each part of a region are satisfied. Relative needs can be expressed as the expected number of hospital admissions in one or more acute clinical specialties that would be generated by an area of residence if national utilization rates by age, sex,and specialty were applied to the local demographic structure. (This is analogous to the method of calculating the patient generating factor in equation 1; it is simply an indicator of *expected* demand.)

Efficiency

The efficiency criterion is defined as choosing a resource configuration that maximizes the benefits to consumers (patients) by satisfying their preferences for treatment in different locations. This criterion is rooted in notions of consumer surplus favored by transport planners, economists, and others, and is presented formally below.

Accessibility

The accessibility criterion is defined in two ways for reasons that will become apparent.

Accessibility (1) - The first is to choose a resource configuration that equalizes the average costs of travel from places of residence to places of treatment. Somewhat related accessibility criteria have precedents particularly in the operations research literature (e.g., Toregas et al. 1971) though very normative assumptions are typically made concerning the nature of demand and the allocation of this demand to particular facilities (for example, the "nearest facility rule"). Here these assumptions are relaxed to preserve the observed spatial choice behavior of patients.

Accessibility (2) - Equalizing the average accessibility costs will be inefficient if the variance in the observed costs between different places of residence is large. Thus a second criterion is defined: it is to choose a resource configuration that minimizes the variance in the accessibility costs from places of residence to places of treatment. In this way, those patients with very high or low accessibility costs may be taken into account.

2.3 Systems Constraints

It is inevitable that in using one or more of these objectives others will conflict in the process. For example, in addition to treating patients, a health care system carries out medical research and trains physicians, nurses, and other personnel. The consequent resource requirements for these activities can conflict with the service requirements of the population (LHPC 1979a). Also, the possibilities for allocating resources among different areas will be constrained by the existing stock of facilities, the availability of land, manpower, economies of scale, finance capital, political, and many other considerations.

These constraints could, if they were sufficiently strong, dominate completely, allowing no room in the strategic plan for any maneuver. In practice, although few new facilities will ever be added to well-established systems and although all the factors described are important to differing degrees, surprisingly large reallocations (for example, -30% to +16% in zones in South-east England between 1975 and 1977) take place through mechanisms such as the updating or enlargement of existing facilities, the closure or reduction in size of old facilities, or a redistribution of more mobile resources such as manpower. The problem, hence, is to include these constraints in a way that will direct the system towards its prime objectives, but with due regard to the operating environment.

Such constraints are clearly important, and it is taken for granted that they would be specified only after detailed discussions with all the actors in the system, including patient representatives, medical, and other experts. Even then, it is anticipated that more than one scenario varying the constraints will be needed to be tested, using the model in a "what if" manner.

3. THE INPUT VARIABLES

There are three input variables in the model--resources, patient generating factors, and accessibility costs--whose estimation is now discussed in more detail before the formal derivation of the model and its application is given.

3.1 Patient Generating Factor

A pgf is calculated as

$$W_i(t) = \sum_k \sum_i P_{ik}(t) u_{mk} \qquad (3)$$

where $P_{ik}(t)$ is the forecasted population in time t, zone i and age-sex category k, and u_{mk} is the projected national hospital utilization rate in clinical specialty m in category k. Although P and u are the dominant considerations in the consumption of health care, the pgf definition is incomplete in the sense that it ignores certain socioeconomic differences among areas that are also believed to influence the use of the services (LHPC 1979a). Some research on identifying these factors has been done and more work is in progress. The projected populations in each area can be determined using conventional demographic methods; a method for forecasting utilization rates is described in LHPC (1979a), LHPC (1979b) and is summarized in Mayhew (1980, Appendix B). The latter assumes a saturation effect, arguing that utilization rates in each clinical category, though generally increasing, will gradually level out in the future.

3.2 Resources

Resources are defined in terms of caseload, the number of patients treated by the system in a particular time period (usually one year). The regional caseload is a function of the availability of hospital beds, the efficiency with which patients can be treated, finance, and other factors. All have to be taken into account. The fundamental relationship in a clinical specialty between cases, beds, and throughput, for example, is

$$B_m(t) = \frac{d_m(t)\,[l_m(t) + t_m(t)]}{365} \qquad (4)$$

where $B_m(t)$ is the number of beds in specialty m in time t, $d_m(t)$ is the number of cases, $l_m(t)$ is the average length of stay between admission and discharge, and $t_m(t)$ is the average length of time between the discharge and admission of a new patient. Lengths of stay depend on clinical practice, the pressure on beds,and other considerations. In some specialties, lengths of stay are declining because of improved methods of treatment, and so it is desirable to introduce these trends into the caseload estimates. Turnover intervals are not constant either, and they must also be carefully considered. Suitable methods for dealing with these measures were used by the LHPC (LHPC 1979a) and are also briefly described in Mayhew (1980).

It is simplest to build the resource measures at a regional level, but if local conditions are quite varied, it may be argued that an aggregation of the separate trends in each place of treatment would be more accurate. In the simpler case only, however,

$$Q(t) = 365 \sum_m \frac{B_m(t)}{d_m(t) + t_m(t)} \qquad (5)$$

where Q is the forecasted caseload to be allocated among the places of treatment. Constraints on each place of treatment

may now be introduced. Suppose that after much analysis, a proportionate increase/decrease of more than ±p in resource levels is regarded as undesirable or unmanageable in a planning period. The constraints are then set as

$$D_j(t)(1 + p) \geq D_j(t) \geq D_j(t)(1 - p) \qquad (6)$$

where D_j is the caseload in j and t is the planning horizon. Between these constraints the system is presumed indifferent to the outcome of the allocative methods.

3.3 Accessibility Costs

Accessibility costs $\{c_{ij}\}$ express the difficulty of someone in zone i being admitted as a patient in treatment zone j. In an HCS the factors determining the way a patient chooses (or is referred to) a particular destination may be complex. In some cases, the decision may be based on convenience; in others it may be the result of a series of referrals from a general practitioner or specialists lower in the HCS hierarchy. In still other cases, the patient may be taken in an emergency to a destination unrelated to his place of residence. In spite of these complexities, a number of measures, including distance, modified distance and journey time have proved reliable indicators of this process, underlining that access is still the dominant consideration in most cases. These measures are further described in Mayhew and Taket (1980).

3.4 Flow Chart

These input variables and the way they are related to the allocation rules are shown in an accompanying flow chart (Figure 3). This provides one example of how the model may be constructed and linked together; it has already been tried in practice but in another context (LHPC 1979a). The outputs are

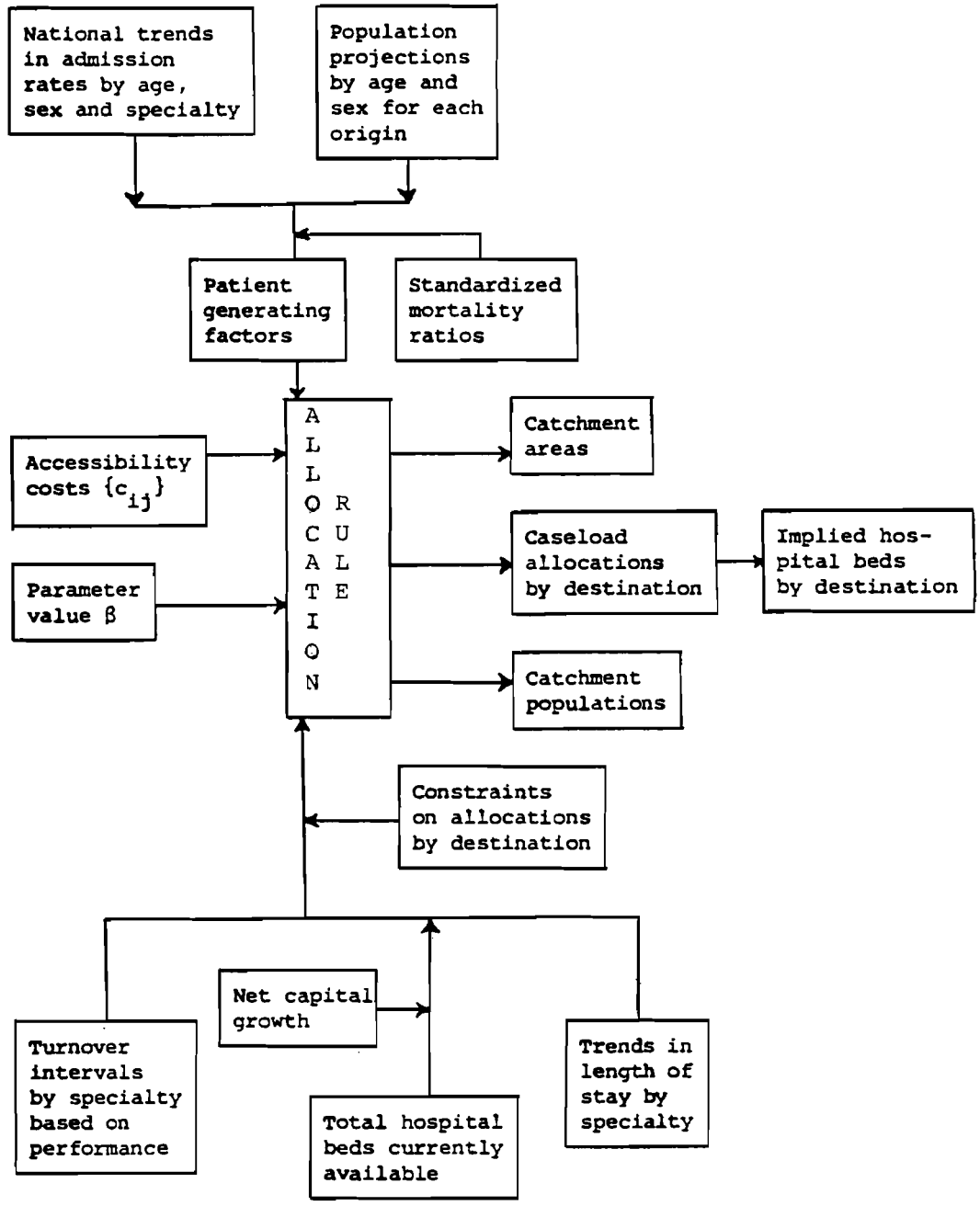

Figure 3. Planning acute in-patient hospital services using the allocation model: the inputs and outputs.

the resources in each place of treatment (right-hand box) and other information of value. These outputs will depend on the total resources available, the configuration of demand, the specification of the constraints, the accessibilities, the model parameter, and the allocation rule. Attention now turns to the formal derivation of the model and the methods for solving it in the case of each allocative criterion.

4. THE MODEL: A FORMAL DERIVATION

It has become customary in recent years to embed gravity models like the one described above (section 3), in types of benefit functions that are derived from concepts of consumer surplus (Wilson and Kirwan 1969; Neuberger 1971; Cochrane 1975; Williams 1977; Coelho and Williams 1978; Leonardi 1980a; Coelho 1980), entropy (Cohen 1961; Wilson 1967; Dacey and Norcliffe 1977; Jefferson and Scott 1979), random utility (Domencich and McFadden 1975; Ben-Akiva and Lerman 1978; Leonardi 1981), or simple utility theory (Mayhew 1981). These provide the models with a consistent theoretical basis, linked to welfare or other considerations. They enable the consideration of a wider range of systems characteristics, while enriching the variety of eventual models and the uses to which they may be put.

The embedding function may be built using only minimal assumptions about the spatial behavior of people, and this is one of their main attractions. In the present case, the function is built for an activity (health care) in which there are excess demands and accessibility costs to pay. The function maximized is subject to the known, and presumed constraints acting in the health care system in order to determine the most likely spatial behavior of the patients.

4.1 Benefit Embedding Functions

This embedding function F is written in a form that incorporates the conclusions of the empirical examples in section 2.1. It takes into explicit consideration the elastic demand mechanism introduced in Leonardi (1980b)

$$F = -\sum_{ij} T_{ij}\left(\log \frac{T_{ij}}{f_{ij}} - 1\right) - \sum_{i} U_i\left(\log \frac{U_i}{h_i} - 1\right) \qquad (7)$$

$$i = \overline{1,I} \quad , \quad j = \overline{1,J}$$

where

T_{ij} = the predicted patient flow between i and j

U_i = unsatisfied demand in i

D_j = caseload capacity in j

f_{ij} = space discount function $e^{-\beta c_{ij}}$, where c_{ij} are the accessibility costs between i and j

β = spatial discount parameter

h_i = a parameter related to the disutility of not receiving treatment

In equation (7), U_i may be thought of as consisting of reported demand in the form of waiting lists, queues, or as unreported demand in the form of sick people who have not presented themselves to a doctor.

Satisfied and unsatisfied demand are related by the identity

$$\sum_{j} T_{ij} + U_i = V_i \qquad (8)$$

where V_i measures the total demand in i.

The problem is to maximize F subject to (8), the total demand in the system, and to a resource constraint in each place of treatment j

$$\sum_i T_{ij} = D_j \tag{9}$$

That is

$$\max_{T,U} [F] \tag{10}$$

This is equivalent to finding the saddle point of the Lagrangian function C, where

$$C = F + \sum_i \lambda_i (V_i - \sum_j T_{ij} - U_i) + \sum_j \nu_j (D_j - \sum_i T_{ij}) \tag{11}$$

and where λ_i and ν_j are the Lagrangian multipliers associated with (8) and (9). The solution is found by equating the first derivatives of C with respect to T_{ij}, U_i, λ_i, ν_j to zero and then solving the $J + I(J + 2)$ equations

$$\frac{\partial C}{\partial T_{ij}} = 0 \tag{12}$$

$$\frac{\partial C}{\partial U_i} = 0 \tag{13}$$

$$\frac{\partial C}{\partial \lambda_i} = 0 \tag{14}$$

$$\frac{\partial C}{\partial \nu_j} = 0 \tag{15}$$

From (11) and (12), and rearranging terms

$$T_{ij} = f_{ij} e^{-(\lambda_i + \nu_j)} \tag{16}$$

Similarly from (11) and (13)

$$U_i = e^{-\lambda_i} h_i \tag{17}$$

Also, from (11), (14), and (16)

$$D_j = \sum_i T_{ij} = e^{-\nu_j} \sum_i e^{-\lambda_i} f_{ij} \tag{18}$$

Therefore

$$e^{-\nu_j} = D_j \left[\sum_i e^{-\lambda_i} f_{ij} \right]^{-1} \tag{19}$$

which in (16) gives

$$T_{ij} = D_j \frac{e^{-\lambda_i} f_{ij}}{\sum_i e^{-\lambda_i} f_{ij}} \tag{20}$$

But, this is

$$T_{ij} = D_j \frac{U_i h_i^{-1} f_{ij}}{\sum_i U_i h_i^{-1} f_{ij}} \tag{21}$$

where $U_i h_i^{-1}$ is the ratio of unsatisfied demand to the disutility of not receiving treatment. Assuming that U_i is sufficiently large so that $\sum_i T_{ij}$ can be considered negligible, U_i from (8) then equals V_i. Defining W_i, the morbidity factor, as $V_i h_i^{-1}$, we obtain the attraction constrained model in equation (1)

$$T_{ij} = \frac{D_j W_i f_{ij}}{\psi_j} \tag{1}$$

where B_j has now been replaced by ψ_j^{-1}

$$\psi_j = \sum_i W_i f_{ij} = B_j^{-1} \tag{22}$$

The path to equation (1) thus makes the nature of the assumptions in the model more clear. We now develop the four criteria (equity, efficiency, accessibility 1 and 2) with which to allocate resources among places of treatment.

4.2 Equity

The objective of the equity criterion is to choose a resource configuration such that the patients generated in each i are in proportion to the relative needs of i.

From (1) and summing over j, the predicted number of patients generated by i is

$$\sum_j T_{ij} = W_i \sum_j \frac{D_j f_{ij}}{\psi_j} \tag{23}$$

since W_i, an index of patient generating potential, is also the expected number of patients, the expression

$$\sum_j \frac{T_{ij}}{W_i} = \sum_j \frac{D_j f_{ij}}{\psi_j} \qquad (24)$$

is therefore the ratio in i of the predicted to the expected. More importantly, it is also the ratio of the predicted service levels to the relative need, and, as we have defined it, the objective is to ensure that this ratio is constant in all origins i by choosing the appropriate values for D_j. However, this quantity cannot be calculated directly without *a priori* knowledge of the service prediction, $\sum_j T_{ij}$. Fortunately, it is completely analogous to base the estimation of this ratio on the total resources available in the system, Q, and W_i. Thus, a new term α is defined which is given by

$$\alpha = \frac{Q}{\sum_i W_i} \qquad (25)$$

This is simply the total resources divided by the total relative needs in the region of interest. If Q reflects resource availability over the whole country, and if the generating factors are based on the expected number of patients, then α will be one. If W_i is calculated in another way this result will not follow automatically.

Taking into account the constraints on change permitted at each destination, the reformulated problem can be written as

$$\underset{D_j}{\text{Min}} \sum_i \left(\sum_j \frac{D_j f_{ij}}{\psi_j} - \alpha \right)^2 = F \qquad (26)$$

subject to

$$D_j(\max) \geq D_j \geq D_j(\min) \qquad \forall j \qquad (27)$$

and

$$\sum_{j \in L} D_j = Q \tag{28}$$

This says: choose D_j to minimize the square of the differences over all origins between the two ratios (Mayhew 1980). The use of the "square" is to eliminate the problems with mixed negative and positive signs. The constraints are on each destination, and they are fixed as appropriate. The total resources, Q, can apply to the whole region, or to a subset L of it. If it is only a subset then the quantity $\sum_i W_i$ should apply over an equivalent subset. Putting

$$\frac{f_{ij}}{\psi_j} = \gamma_{ij} \tag{29}$$

expanding (26), and ignoring the constant term $I\alpha^2$, where I is the number of origins, we obtain

$$F = \frac{1}{2} \underline{D}^T A \underline{D} - \underline{b}^T \underline{D} \tag{30}$$

where $\underline{D}$ and $\underline{D}^T$ is the vector of resources and its transpose

$$\underline{D} = \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_j \\ \vdots \\ D_n \end{bmatrix} \quad \text{and} \quad \underline{D}^T = [D_1 \; D_2, \ldots, D_j, \ldots, D_n] \tag{31}$$

$$n = J$$

A is a symmetric matrix composed of the following elements

$$A = \begin{bmatrix} 2\sum_i \gamma_{i1}^2 & 2\sum_i \gamma_{i1}\gamma_{i2} & \cdots & 2\sum_i \gamma_{i1}\gamma_{in} \\ 2\sum_i \gamma_{i2}\gamma_{i1} & 2\sum_i \gamma_{i2}^2 & \cdots & 2\sum_i \gamma_{i2}\gamma_{in} \\ \vdots & & & \vdots \\ 2\sum_i \gamma_{ij}\gamma_{i1} & \cdots & 2\sum_i \gamma_{ij}^2 & 2\sum_i \gamma_{ij}\gamma_{in} \\ \vdots & & & \vdots \\ 2\sum_i \gamma_{in}\gamma_{i1} & \cdots & & 2\sum_i \gamma_{in}^2 \end{bmatrix} = \{a_{ij}\} \quad (32)$$

$\underline{b}^T$ is the transpose of the vector $\underline{b}$ in which the elements are

$$\underline{b} = \begin{bmatrix} 2\alpha \sum_i \gamma_{i1} \\ 2\alpha \sum_i \gamma_{i2} \\ \vdots \\ 2\alpha \sum_i \gamma_{ij} \\ \vdots \\ 2\alpha \sum_i \gamma_{in} \end{bmatrix} = \{b_j\} \quad (33)$$

Similarly (27) and (28) can be written in matrix notation

$$\underline{D}_{min} \geq \underline{D} \geq \underline{D}_{max} \quad (34)$$

and

$$\underline{C}^T\underline{D} = Q \tag{35}$$

where $\underline{C}^T$ is a 1 × n vector transpose with all the elements set equal to one. Equations (26), (27), and (28) have now been put into the standard form expected by a general quadratic programming algorithm. The matrix A is always positive definite or semi-definite indicating that global minima are obtainable. In an unconstrained problem the minimum of F is found when the vector of first derivatives disappears. That is when

$$g = \nabla(\frac{1}{2}\ \underline{D}^T A\underline{D} - \underline{b}^T\underline{D}) = A\underline{D} - \underline{b} = 0 \tag{36}$$

Details of the solution method for this problem with and without constraints are contained in Fletcher (1970, 1971) and briefly in Mayhew (1980).

The equity problem, it should be noted, also has an interesting counterpart. Instead of redistributing the resources between each place of treatment j, the same equitable result may be attained by levying an "accessibility tax" on each place of residence i to regulate demand. While such a tax would almost certainly be unpopular, it is of theoretical value since it illustrates the symmetry of the allocation problem. The derivation of the tax and its interpretation are shown in the Appendix.

4.3 Efficiency

Under the efficiency criterion the objective is to allocate D, so that patient preferences for places of treatment are maximized. These preferences are subject to the same constraints as applied in the equity case, that is on each place of treatment and on the total resources available, Q. Putting equation (1) in (7), summing over i, and ignoring terms two and three, which become constants, it is found

$$F = - \sum_j D_j \left(\log \frac{D_j}{\psi_j} - 1 \right) \tag{37}$$

where 1 in (37) replaces another constant without loss of generality. The reformulated problem becomes, therefore,

$$\underset{D_j}{\text{Max}} \; [F] \tag{38}$$

subject again to

$$D_{j(max)} \geq D_j \geq D_{j(min)} \tag{27}$$

and

$$\sum_{j \epsilon L} D_j = Q \tag{28}$$

This is equivalent to finding the saddle point of the Lagrangian function H where

$$H = F + \lambda(Q - \sum_j D_j) + \sum_j \mu_j [D_{j(max)} - D_j] - \sum_j \eta_j [D_{j(min)} - D_j] \tag{39}$$

and where λ, μ_j and η_j are the Lagrange multipliers associated with the resources available, Q,and the inequality constraints in (27). The solution to this maximization problem is found by solving the 3J + 1 equations

$$\frac{\partial H}{\partial D_j} = 0 \tag{40}$$

$$\frac{\partial H}{\partial \lambda} = 0 \tag{41}$$

$$\frac{\partial H}{\partial \eta_j} = 0 \tag{42}$$

and

$$\frac{\partial H}{\partial \mu_j} = 0 \tag{43}$$

plus the complementarity slackness conditions

$$\mu_j [D_{j(max)} - D_j] = 0 \tag{44}$$

$$\mu_j, \eta_j \geq 0$$

$$\eta_j [D_{j(min)} - D_j] = 0 \tag{45}$$

It is easily shown that H is optimal when

$$D_j = \psi_j \, e^{(\eta_j - \mu_j - \lambda)} \tag{46}$$

But, from (28)

$$Q = \sum_j D_j = e^{-\lambda} \sum_j \psi_j \, e^{-\mu_j + \eta_j} \tag{47}$$

Making $e^{-\lambda}$ the subject of (47) and substituting in (46), the result arises

$$D_j = Q \, \frac{\psi_j \, e^{-\mu_j + \eta_j}}{\sum_j \psi_j \, e^{-\mu_j + \eta_j}} \tag{48}$$

In the case when there are no bounds on D_j operating [see equation (27)], (46) becomes

$$D_j = Q \frac{\psi_j}{\sum_j \psi_j} \qquad Q \geq D_j > 0 \tag{49}$$

since

$$\mu_j = \eta_j = 0 \tag{50}$$

Equation (49) is the basic allocation formula that matches the resources in j with patient preferences for treatment in that location. The preference term is ψ, which is the sum of the pgfs discounted by the accessibility costs [equation (22)]. It is a measure of the total demand potential on j after accessibility costs have been paid. Thus, the resources are divided between places of treatment simply by proportioning Q according to potential on j divided by the sum of all the potentials on all j.

4.4 Accessibility (1)

The average accessibility costs from i to all j is defined as

$$c_i = \frac{\sum_j T_{ij} c_{ij}}{\sum_j T_{ij}} = \frac{\sum_j \frac{D_j f_{ij} c_{ij}}{\psi_j}}{\sum_j \frac{D_j f_{ij}}{\psi_j}} \tag{51}$$

Since the criterion requires that c_i be constant, it may be replaced by $\bar{c}$, where $\bar{c}$ is either presumed beforehand or it is based on the current system's average, that is

$$\bar{c} = \frac{\sum_i \sum_j T_{ij} c_{ij}}{\sum_i \sum_j T_{ij}} \tag{52}$$

The objective may now be defined. It is

$$\min_{D_j} [G] \tag{53}$$

subject to

$$D_{j(max)} \geq D_j \geq D_{j(min)} \tag{27}$$

and

$$\sum_{j \in L} D_j = Q \tag{28}$$

where

$$G = \sum_i (c_i - \bar{c})^2 \tag{54}$$

This says: minimize the differences in all i between the average accessibility costs to j and a supplied average, $\bar{c}$, subject to the usual constraints. Equation (54) has an interesting property; it is a homogeneous function of degree zero. Hence, the following property holds

$$G(k\underline{D}) = \sum_i [c_i(k\underline{D}) - \bar{c}]^2 = \sum_i [c_i(\underline{D}) - \bar{c}]^2 = G(D) \tag{55}$$

where k is a constant ($\neq 0$) and $\underline{D}$ is a vector with J elements. Equation (55) describes a lined surface in J dimensions with the lines having directional cosines proportional to $\underline{D} = (D_1, \ldots, D_n)$. Along any line the average cost, and hence G, is unchanged for different values of D, thus indicating an infinite number of solutions to this problem. However, providing the resource constraint in equation (28) is applied, the problem has a well-defined solution.

4.5 Accessibility (2)

The variance criterion is constructed in a similar way. The variance in the travel costs from i to all j is defined as

$$v_i = \frac{\sum_j T_{ij}(c_{ij}-\bar{c})^2}{\sum_j T_{ij}} = \frac{\sum_j \frac{D_j f_{ij}}{\psi_j}(c_{ij}-\bar{c})^2}{\sum_j \frac{D_j f_{ij}}{\psi_j}} \qquad (56)$$

The objective is then written

$$\min_{D_j} [S] \qquad (57)$$

subject to

$$D_{j(max)} \geq D_j \geq D_{j(min)} \qquad (27)$$

and

$$\sum_{j \in L} D_j = Q \qquad (28)$$

where

$$S = \sum_i v_i \qquad (58)$$

Like the first accessibility criterion, the second is also homogeneous of degree zero, the objective function describing again a lined surface in J dimensions.

The 2-origin 2-destination Problem

Figure 4 shows sketches of all four criteria in the simplest of possible systems: two origins and two destinations. On the axes in the plane are D_1 and D_2, the two unknowns. On the vertical axis in arbitrary units are the values of the four objective functions. The regional resource constraint is represented by the diagonal AB along which $D_1 + D_2 = Q$. The desired values of D_1 and D_2 are located on AB at the maximum or minimum of the respective functions. When upper and lower bounds on D_j are applied, the plane is divided by vertical and horizontal lines into a feasible and an infeasible region; the optimum value on each criterion is still lying on AB but inside in the feasible part. Figure 4 also shows the important result that each criterion selects in general a different set of resource allocations from the others, thus drawing attention to their incompatability. To determine the suitability of these criteria, the results of the application to a planning problem in the United Kingdom are now described.

5. APPLICATION

The above-described methods have been applied and tested on 1977 data for the London region in England. London forms a particularly appropriate application since it has especially severe planning problems that have resisted solution by more conventional approaches. Approximately 7 million people live in the area covered, and it is served by about 200 hospitals treating approximately 1 million in-patient cases each year. Because of changes in the size and demographic structure of the population, health authorities are interested to know which facilities to enlarge, reduce in size, or close altogether. The existing pattern of patient flows between areas, however, is complex: this is due to the proximity of facilities (particularly the relative over-concentration in the city center), the ready availability of transport services, and other factors. In addition, there are constraints on change that are imposed

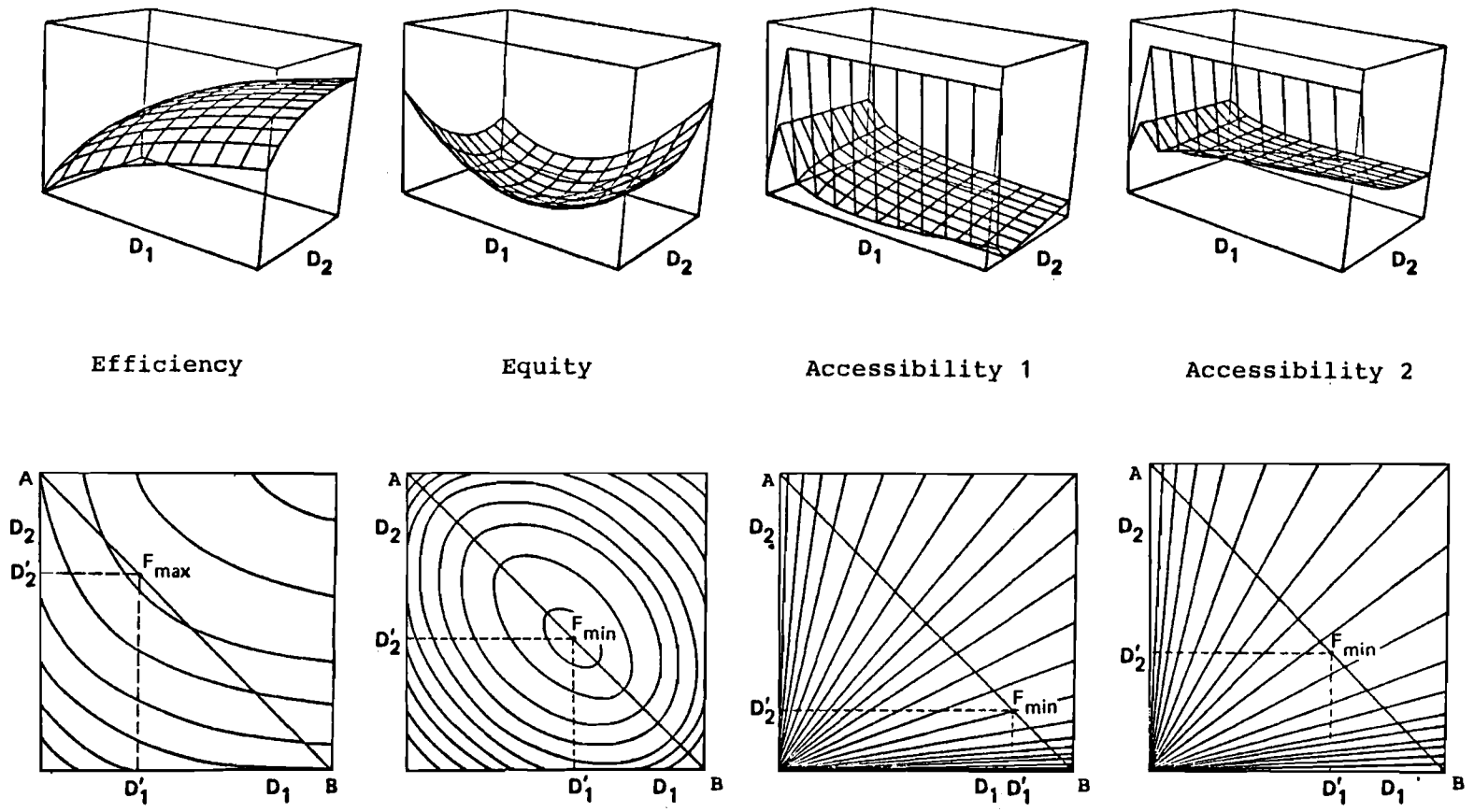

Figure 4. Three-dimensional representations and maps of the objective functions for the 2-origin, 2-destination case.

by the condition of the existing hospital stock, the availability of land, financing, and other resources. Finally, London is a national and international center for medical education and research whose activities in these fields must be taken into account in the resource allocation process. To these specific factors must be added the differential trends in treatment that are changing the patient mix and type of care received, with important implications for hospital throughput and hence case-load capacities.

5.1 Zoning System

In Figure 5 two maps show the 33 origin zones (administrative boroughs of the Greater London Council, GLC) and 36 destination zones (Health Districts) used in these applications. The names of these zones may be found in Table 1 in Mayhew (1980:24). In addition to these, there is one external zone to close the system. The model for this region was constructed from an aggregate of 23 acute specialties, a list of which is shown in Table 1 of Mayhew and Taket (1980:16). Details of the calibration procedure are also found in this reference, while the results of validation tests to check the predictive capability of the model are given in Mayhew and Taket (1981). Here, all that is essential, in addition to the input data, is a value for the β parameter in equation (1), which was obtained from the above work. It is 0.367.

5.2 Presentation of Outputs

The most convenient ways of illustrating the outputs of these procedures are with bar charts, showing the proportionate changes in allocations, and scatterdiagrams. Scatterdiagrams show the relationships--both before and after the application of the methods--between the numbers of patients generated in i,

A) Origin zones

B) Destination zones

Figure 5. The Greater London Council: definition of zones.

an origin zone (i.e., $\sum_j T_{ij}$), and the relative need in i scaled by α--the regional service-demand ratio given in equation (25) (i.e., αW_i). A linear equation fitted to this scatter will thus give the extent to which the reallocation process has satisfied the relative needs of the population. In the equity case only the result should be an equation with a slope coefficient $\hat{b}$ equal to one and an intercept term $\hat{a}$ that is not statistically significantly different from zero. When the coefficient of explanation R^2 is also one, it means the equity criterion has been met exactly. In practice, the value of R^2 is reduced according to the stringency of the constraints applied on the destinations, $D_{j(min)}$ and $D_{j(max)}$. For the other cases, the properties of the resultant scatters are completely different, but as will be seen, they usually provide sufficient information to judge the effectiveness of each criterion. (A straight line in the efficiency case is also obtained when D_j is plotted on $\gamma\psi_j$, where $\gamma = Q/\sum_j \psi_j$. This would be an alternative way of presenting the results.)

5.3 Tests

Each criterion has been thoroughly tested using the existing and hypothetical data to represent both the current situation and possible development scenarios (changes in supply and demand). Some of these scenarios were deliberately exaggerated to see how the methods performed when they were stretched for particular input sets. Only the results obtained with the current data sets are reported, although all the developmental runs of the methods have been taken into account. For simplicity and brevity, only two tests are shown: one with a lower bound each on destination and one without. That is

<u>Test 1:</u> $\quad Q \geq D_j \geq D_j(1 - 0.25)$

<u>Test 2:</u> $\quad Q \geq D_j \geq 0$

The upper bound in test 1 has been left open (although Q, of course, is the maximum that can be allocated) to see where the major shortfalls in resources are predicted to occur; the lower bound has been arbitrarily fixed to 75% of the current value. In test 2 the lower bound is simply zero to avoid negative allocations.

5.4 Allocative Behavior

Figures 6 and 7 show the predicted percentage change in allocations for each test. In test 1, the influence of the 25% lower bound shows up strongly in the negative part of the charts, whereas in test 2 it is seen that the allocations can give extreme solutions, emphasizing allocations to only one or two locations. In the experiments carried out, the equity criterion is always the least susceptible to this behavior, whereas efficiency and accessibility are the most susceptible. In the efficiency case, for example, the results are especially sensitive to the measurement of the local accessibility costs; the reasons for the very unusual large allocations in test 2 to zones 14 and 23 by accessibility 1 are unclear, however. It was generally found that the spatial pattern of reallocations are more intuitive in the cases of equity and efficiency than for accessibility tests 1 and 2, and this empirical feature makes them more attractive as allocative criteria. For example, the charts in both test 1 and 2 show that the equity and efficiency criteria tend to peripheralize the available resources to zones lying closer to the perimeter of the urban region. This is consistent with other findings (e.g., LHPC 1979a) that show the central area is relatively over-provided with resources.

5.5 Patient Behavior

The effects in these reallocations on the service levels (numbers treated) of the population in each place of residence i is shown in Figure 8. (Figure 9 shows the existing service levels plotted on relative needs.) As is seen in Figure 8 (a and b)

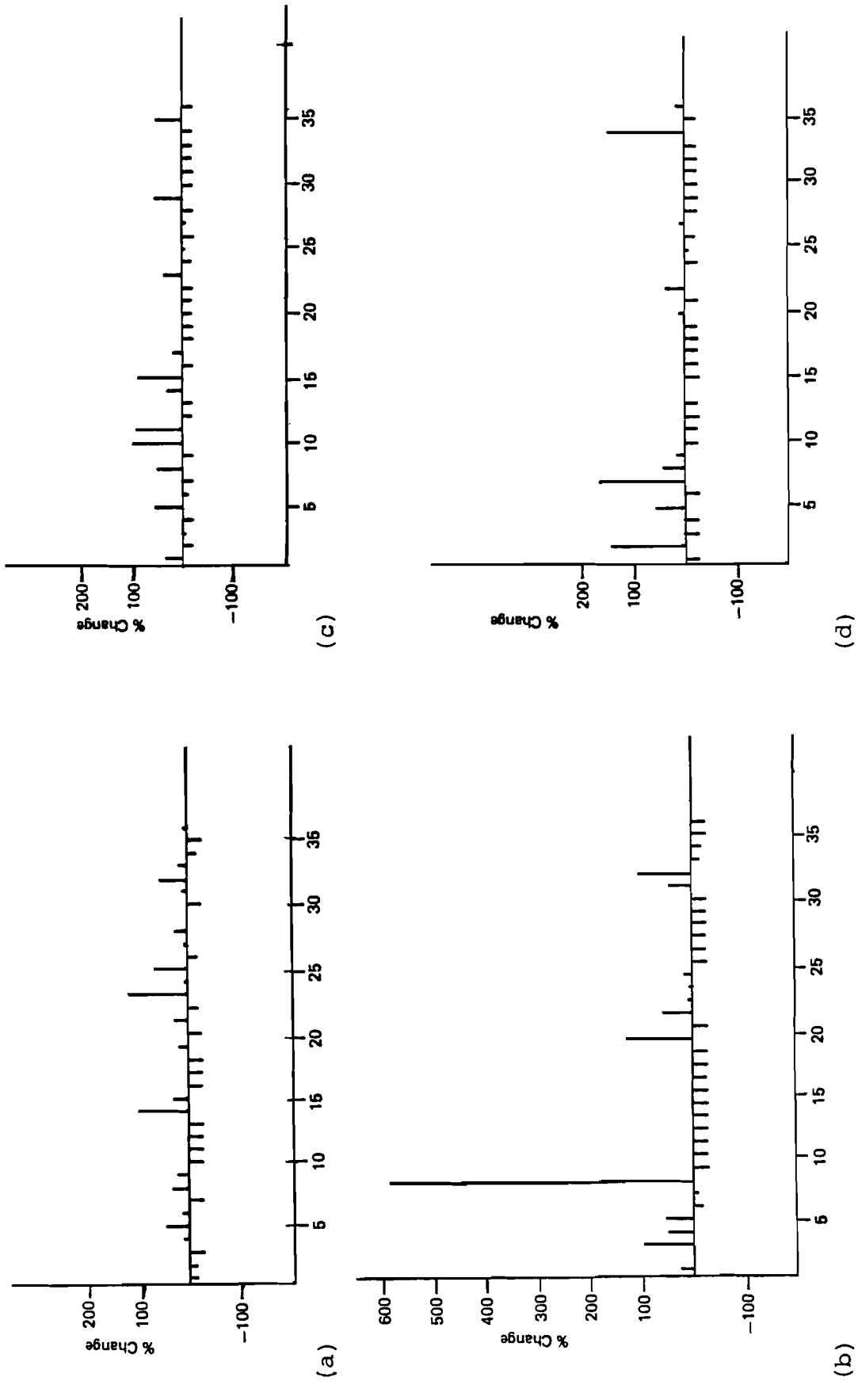

Figure 6. Percentage changes in resource allocations to destination zones under test 1. [(a) equity, (b) efficiency, (c) accessibility 1, (d) accessibility 2]

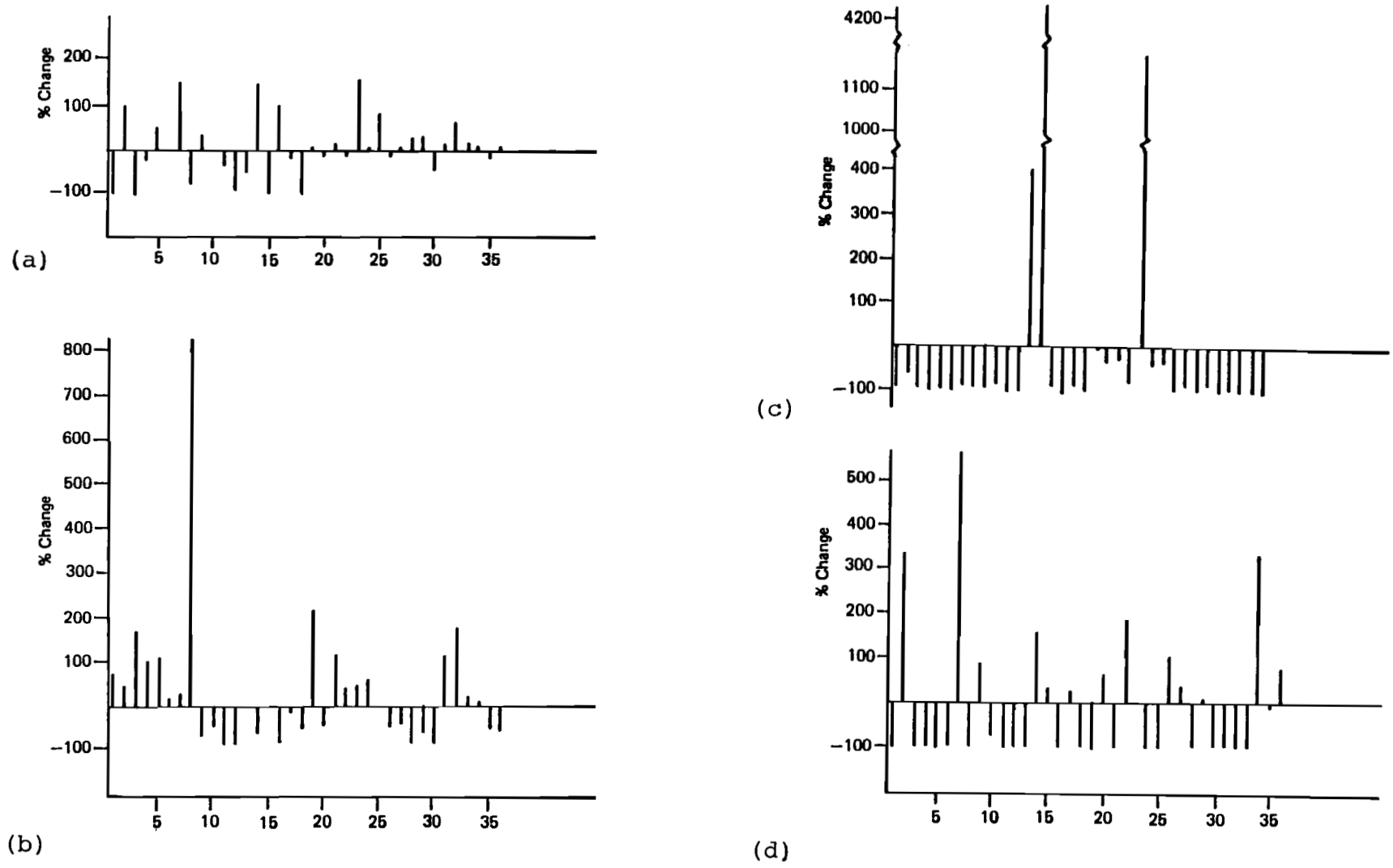

Figure 7. Percentage changes in resource allocations to destination zones under test 2. [(a), equity, (b) efficiency, (c) accessibility 1, (d) accessibility 2]

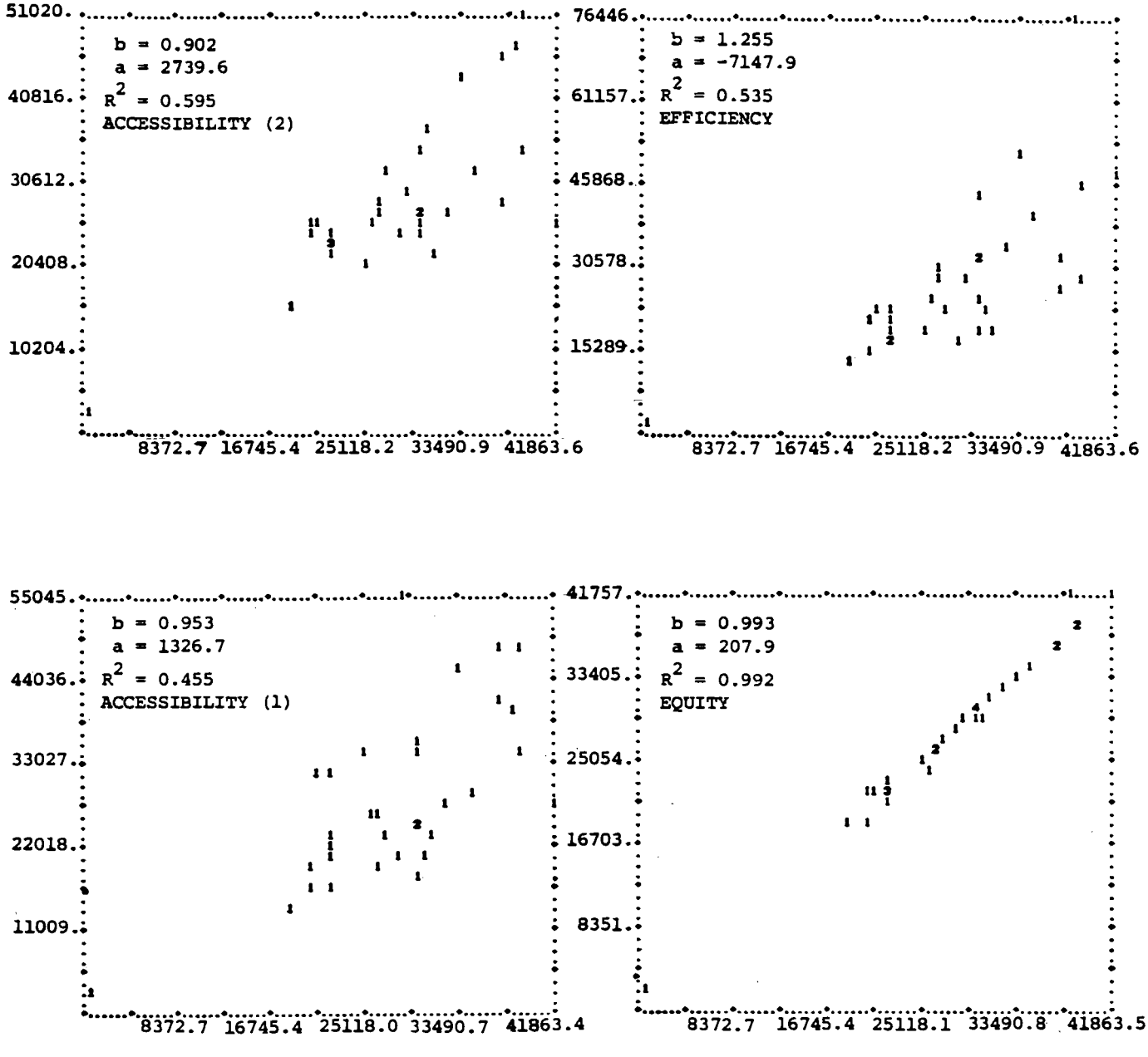

Figure 8a. Plot of patients generated in i on the relative needs of i (test 1).

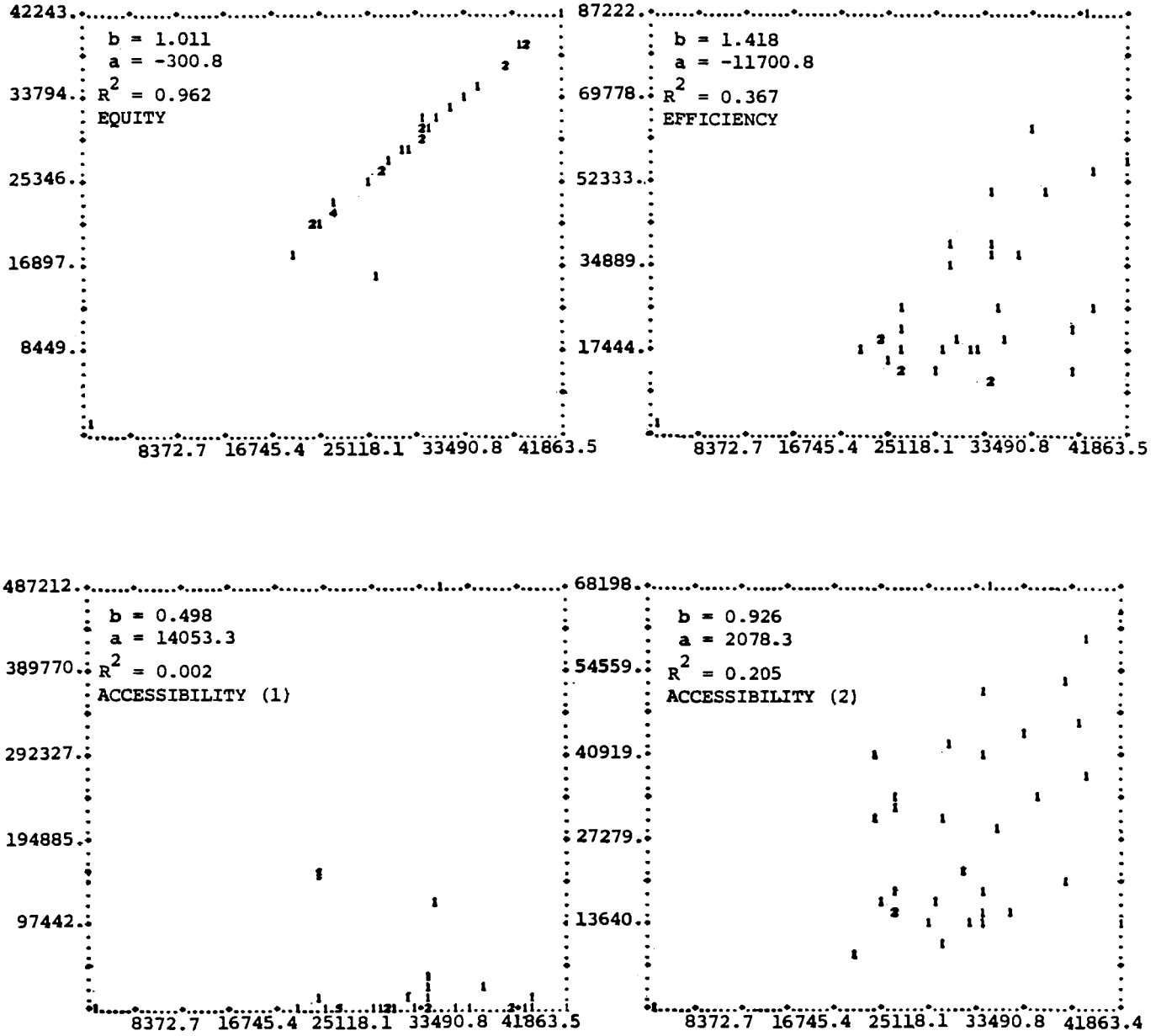

Figure 8b. Plot of patients generated in i on the relative needs of i (test 2).

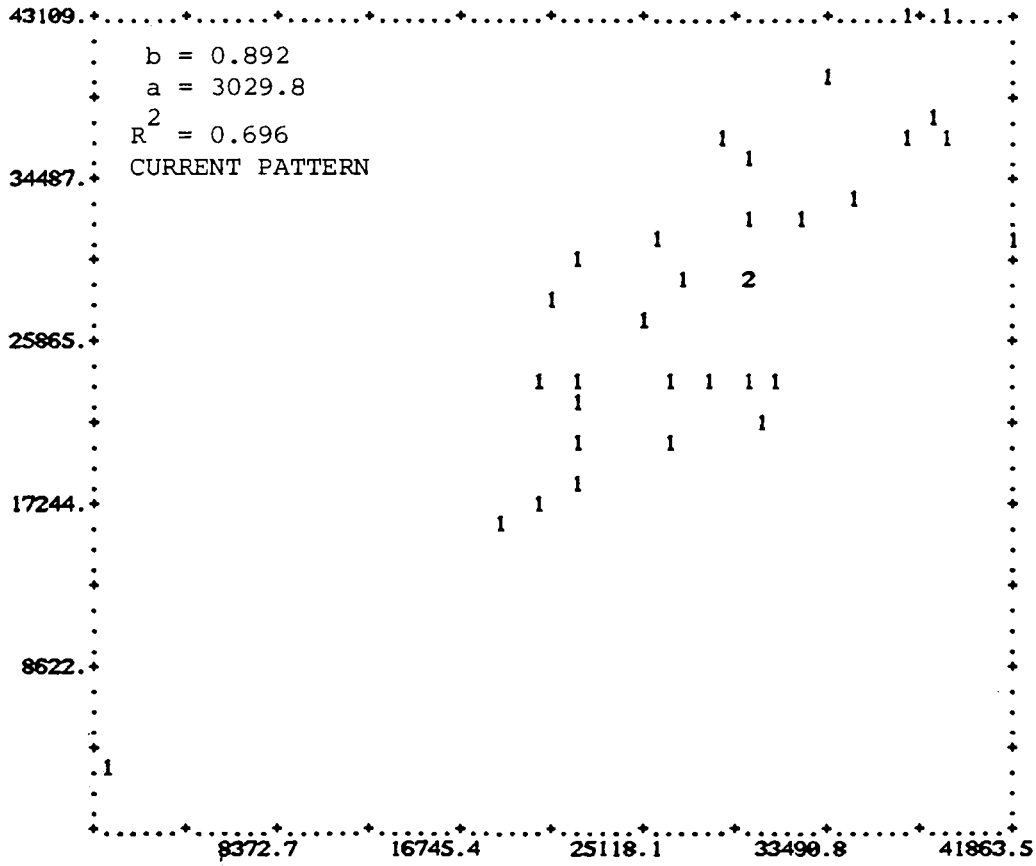

Figure 9. Plot of existing numbers of patients generated in i on the relative needs of i.

the equity criterion reproduces the straight line as desired in both tests. An encouraging feature in all the experiments is the stability of the slope and intercept terms (which is necessary under the equity definition) even during some very severe tests of the method. Furthermore, it was found that large gains in equity were attainable even when the constraints in change were very tight (say D_j ±5%). In test 2, the unconstrained case (Figure 8b), an outlier among the data points is observed for the equity case: fortunately this behavior never arises in more realistic applications that use constraints.

The other criteria do not have the above slope property, and the values of R^2 they give are, as is seen in Figure 8, always less than in the equity case for the same sets of constraints. This underlines the fact that equity, efficiency and accessibility (1 and 2) are incompatible goals in that it is impossible with this data and this model to achieve all four simultaneously.

The effects of the unusual allocations on service levels by accessibility 1 found in test 2 (see Figure 7) is shown in Figure 8b. The result is clearly unsatisfactory in that, as is shown, no attempt is made to reconcile the resources allocated with the relative needs of the population ($R^2 = 0.002$). On this basis and on the basis of other experiments, it thus seems unreasonable to proceed with this criterion. The case for rejecting accessibility 2, however, is much less clearcut. The main problems with it seem to be firstly its somewhat unpredictable behavior in sensitivity tests carried out on the constant c in equation (58), and secondly the often counter-intuitive results obtained. These make it difficult to understand the precise mechanisms of this method. Nevertheless, further applications are needed to settle these points.

Sensitivity Analysis

The equity and efficiency cases were thus selected for further sensitivity analsyses. This involves an unconstrained model of the type used in test 2 but in which the β parameter is allowed to vary over a wide range. Although in practice this parameter is expected to change very little, the experiment is necessary to test the logic of the allocations when the criteria are exposed to extremes of behavior. For instance, a value of β equal to zero implies that there are no accessibility costs to pay, whereas a large value implies very large costs and therefore a high space discount premium. Tables (1) and (3) indicate facility behavior in each treatment district for different β values. A dot indicates that all the facilities in a district have been closed. Tables (2) and (4) show the regression coefficients and values for R^2.

Table 1. Sensitivity of facility behavior with respect to β: the equity case. (Zones where all facilities have been closed are indicated by black dots.)

Zone j \ β	0.005	0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	1.0	3.0	5.0	8.0
1	●	●	●	●	●	●	●		●							●
2	●	●	●	●	●										●	
3	●	●	●	●	●	●			●	●	●	●				
4	●	●	●													
5	●	●	●	●												
6	●	●	●													
7	●	●	●	●	●	●	●	●					●	●	●	
8	●	●	●	●	●	●	●	●								
9	●	●														
10	●	●	●	●												
11	●	●														●
12	●									●	●	●				
13	●	●	●	●	●											
14	●	●														
15	●	●	●				●	●	●	●		●	●	●	●	●
16	●	●	●													
17	●	●	●													
18						●	●	●	●	●	●	●	●	●	●	
19	●	●	●	●	●	●										
20	●	●														
21	●	●	●	●												
22	●	●	●		●											
23	●	●	●	●	●											
24	●	●	●	●												
25	●	●	●	●												
26	●	●														
27	●	●	●													
28	●	●	●	●												
29	●	●				●	●	●								
30	●	●								●	●	●	●	●	●	●
31	●	●														
32	●	●	●													
33	●	●	●													
34	●	●	●	●	●									●	●	
35	●	●	●	●	●	●										
36	●	●														

Table 2. Sensitivity analysis of β: regression results for the equity case.

Regression Results	β 0.005	0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	1.0	3.0	5.0	8.0
R^2	1.00	0.99	0.96	0.95	0.96	0.98	0.98	0.98	0.9	0.95	0.93	0.91	0.76	0.67	0.67	0.67
$\hat{b}$	0.64	0.64	1.06	1.07	1.06	1.03	1.02	1.00	1.01	1.02	1.03	1.04	1.16	1.28	1.29	1.30
$\hat{a}$	-18.2	-3.15	-1734	-2018	-1552	-884	-424	-122	-256	-427	-698	-1022	-4367	-7895	-8158	-8385
Total zones with open facilities	1	2	11	20	25	28	30	31	32	31	32	31	32	31	30	32

KEY

R^2 = coefficient of explanation

$\hat{b}$ = slope

$\hat{a}$ = intercept

Table 3. Sensitivity of facility behavior with respect to β: the efficiency case.

Zone \ β	<1	1.5	1.75	2.0	2.25	2.5	2.75	3.0	3.5	4.0	4.5	5.0	5.5	6.0	7.0	8.0
1												•	•	•	•	•
2									•	•	•	•	•	•	•	•
3																
4																
5																
6														•	•	•
7					•	•	•	•	•	•	•	•	•	•	•	•
8																
9							•	•	•	•	•	•	•	•	•	•
10								•	•	•	•	•	•	•	•	•
11		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
12			•	•	•	•	•	•	•	•	•	•	•	•	•	•
13											•	•	•	•	•	•
14				•	•	•	•	•	•	•	•	•	•	•	•	•
15									•	•	•	•	•	•	•	•
16		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
17										•	•	•	•	•	•	•
18																
19																
20												•	•	•	•	•
21																
22													•	•	•	•
23									•	•	•	•	•	•	•	•
24																
25									•	•	•	•	•	•	•	•
26										•	•	•	•	•	•	•
27										•	•	•	•	•	•	•
28		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
29						•	•	•	•	•	•	•	•	•	•	•
30			•	•	•	•	•	•	•	•	•	•	•	•	•	•
31																•
32																
33																•
34						•	•	•	•	•	•	•	•	•	•	•
35								•	•	•	•	•	•	•	•	•
36									•	•	•	•	•	•	•	•

Table 4. Sensitivity analysis of β: regression results for the efficiency case.

Regression Results	β <1	1.5	1.75	2.0	2.25	2.5	2.75	3.0	3.5	4.0	4.5	5.0	5.5	6.0	7.0	8.0
R^2	0.17	0.13	0.12	0.11	0.10	0.09	0.09	0.08	0.07	0.07	0.06	0.06	0.06	0.05	0.51	0.05
$\hat{b}$	2.40	2.87	3.00	3.10	3.17	3.21	3.23	3.25	3.25	3.24	3.22	3.20	3.18	3.17	3.16	3.15
$\hat{a}$	-39×10^3	-52×10^3	-56×10^3	-58×10^3	-60×10^3	-61×10^3	-62×10^3	-63×10^3	-62×10^3	-62×10^3	-62×10^3	-61×10^3	-61×10^3	-61×10^3	-60×10^3	-60×10^3
Total zones with open facilities*	36	33	31	30	29	27	26	24	19	16	15	14	12	11	11	9

*Allocations of $D_j \leq 0$ are impossible with the efficiency criterion [equation (47)]. Thus a "closed" facility is said to occur when $D_j < 10$.

(a) Equity

For $\beta = 0.005$ the only facilities open are at the city center itself (zone 18). This seems most logical as this zone is a focus for the whole region. The first facilities in outer zones appear when $\beta = 0.1$. When $\beta = 0.2$, the facilities in the center close because as costs get higher, needs are better served locally rather than centrally. As β increases further, more suburban facilities open until a maximum of 32 out of 36 zones have resources allocated to them. The special case when $\beta = 0$ should also be noted (i.e., no accessibility costs at all). From equations (22) and (26), we see that the coefficients γ_{ij} become constant and that the objective function reduces to

$$F = \sum_i \left(\frac{\sum_j D_j}{\sum_i W_i} - \alpha \right)^2 \tag{59}$$

Since $\sum_j D_j = Q$ and since $\alpha = Q/\sum_i W_i$, F will be a minimum no matter how the resources are allocated. Thus there are an infinite number of equitable solutions to this case.

(b) Efficiency

Facility behavior under the efficiency criterion is the opposite of equity. When β is zero, equation (47) reduces to

$$D_j = \frac{Q}{J} \tag{60}$$

where J equals the number of treatment zones. Thus each district receives an identical one J-th share of the available resources Q. As β increases, the more accessible locations to demand (i.e., those with high potentials ψ_j) begin to dominate the solution, so that gradually the zones with less potential become ignored and the facilities in them are closed. Another major

difference with the equity solution is that the central facilities (zone 18) always remain open, whereas in the equity case they are closed ($0.2 \leq \beta \leq 8.0$).

6. THE EQUITY-EFFICIENCY TRADE-OFF MODEL

In view of the different resource configurations produced by the equity and efficiency criteria, it seems reasonable for certain types of health care systems to design a model that permits the user to trade off one goal against the other. In order to analyze these trade-offs the following mathematical programming problem is constructed

$$\max_{D} F(D) = \theta V_1(D) + (1 - \theta) V_2(D) \tag{61}$$

subject to

$$\sum_{j \in L} D_j = Q \tag{28}$$

$$D_{j(\max)} \geq D_j \geq D_{j(\min)} \tag{27}$$

where $D = \{D_j\}$, $j = \overline{1,J}$

$$V_1(D) = - \sum_j D_j \left(\log \frac{D_j}{\psi_j} - 1 \right) \tag{37}$$

$$V_2(D) = - \sum_i \left(\sum_j \frac{D_j f_{ij}}{\psi_j} - \alpha \right)^2 \tag{26}$$

and θ is a trade-off parameter. Equation (61) is a mixture of the equity and efficiency objective functions. It is to be maximized subject to the usual constraints in (27) and (28).

This is a concave programming problem with simple linear constraints. A well-known method to solve it is the Frank-Wolfe Method (Frank and Wolfe 1956), which in this case takes a simple form. The iterations of the method are based on using a linear approximation to equation (61) in order to find best directions of increase. The linear sub-problem for (61), (27), and (28) is written

$$\max_{D} \sum_{j} D_j F'(D^0) \tag{62}$$

where D^0 is the best guess solution so far and $F'(D^0)$ are the derivatives evaluated at the point D^0.

This is derived by expanding $F(D)$ in a Taylor expansion around D^0, truncated to the first-order terms. These terms desribe the tangent plane to (61), and if the constant terms are ignored the result simplifies to (62). Sub-problem (62) is now a simple continuous knapsack problem, which is easily solved for this special case (e.g., see McMillan 1975).

The solution to (62), (27), and (28) is now used to determine the best direction for an improvement in (61). That is

$$d = D^* - D^0 \tag{63}$$

where D^* is the solution just obtained, (63). The best guess solution to (61), (27), and (28) is now found by solving the univariate maximization problem

$$\max_{0 \le \lambda \le 1} F(D^0 + \lambda d) \tag{64}$$

This is given by

$$D^1 = D^0 + \lambda d \tag{65}$$

Problem (64) can be solved, for instance, by the Newton-Raphson method. These steps, sub-problems (62), (21), (28), and (64), may then be repeated until convergence. The method is usually fast in the first few iterations, although it is difficult to reach a much higher level of precision in further steps. However, it is well suited to the type of sensitivity analysis required in the trade-off model whose application is described below.

6.1 Trade-off Results

Figure 10 shows the results for the service levels in the origin zones based on different values of the trade-off parameter (TOP), which range from pure efficiency (TOP = 1.0) to pure equity (TOP = 0.0). No constraints, only $D_j \geq 0$, have been applied in this example, although the algorithm developed has the capability of incorporating constraints. As is seen, by reducing the effect of the efficiency component, the scatter of points gradually assumes the characteristic straight-line form with a slope b becoming closer to 1.0. Note that the trade-off parameter must first be very small ($< 0.5 \times 10^{-5}$) before the equity criterion takes effect. This is simply a reflection of the different ways the individual functions are constructed and their component values. The general form of the trade-off curve is shown in Figure 11. Since each part of the function is measured in different units and since each has a range of values dependent on the input variables, it was found useful to standardize the axes in this figure in the range 0-100.

The result is the smooth curve in Figure 11, points on which indicate the indexed values (0-100) of the component functions. Unfortunately, it is not possible to talk in terms of an allocation, which expresses the result as percent efficiency and percent equity. The main advantage of this approach is to allow a user to test a wider range of planning options that are not

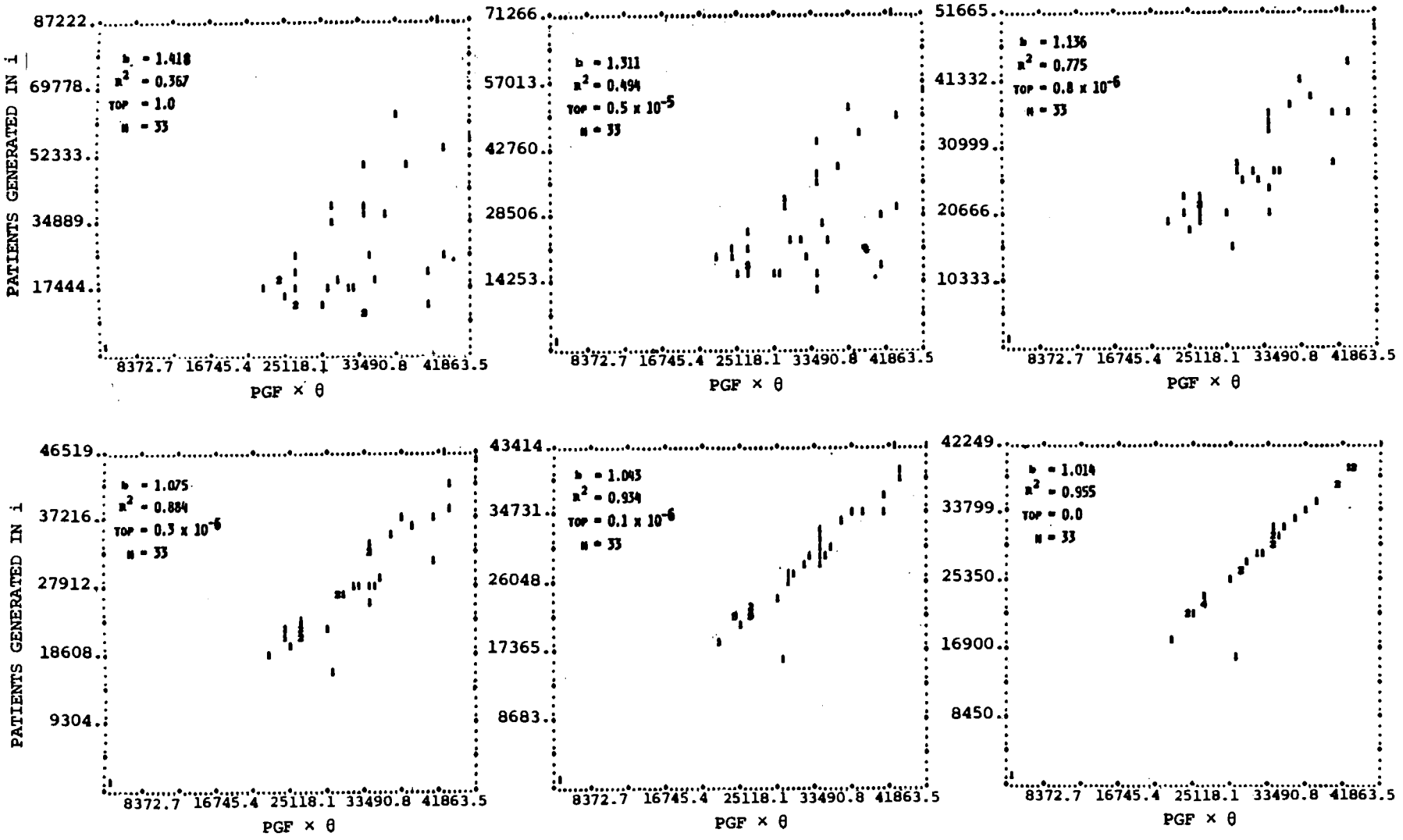

Figure 10. Results for the trade-off model for different values of the trade-off parameter: a plot of predicted patients generated in i on relative needs of i.

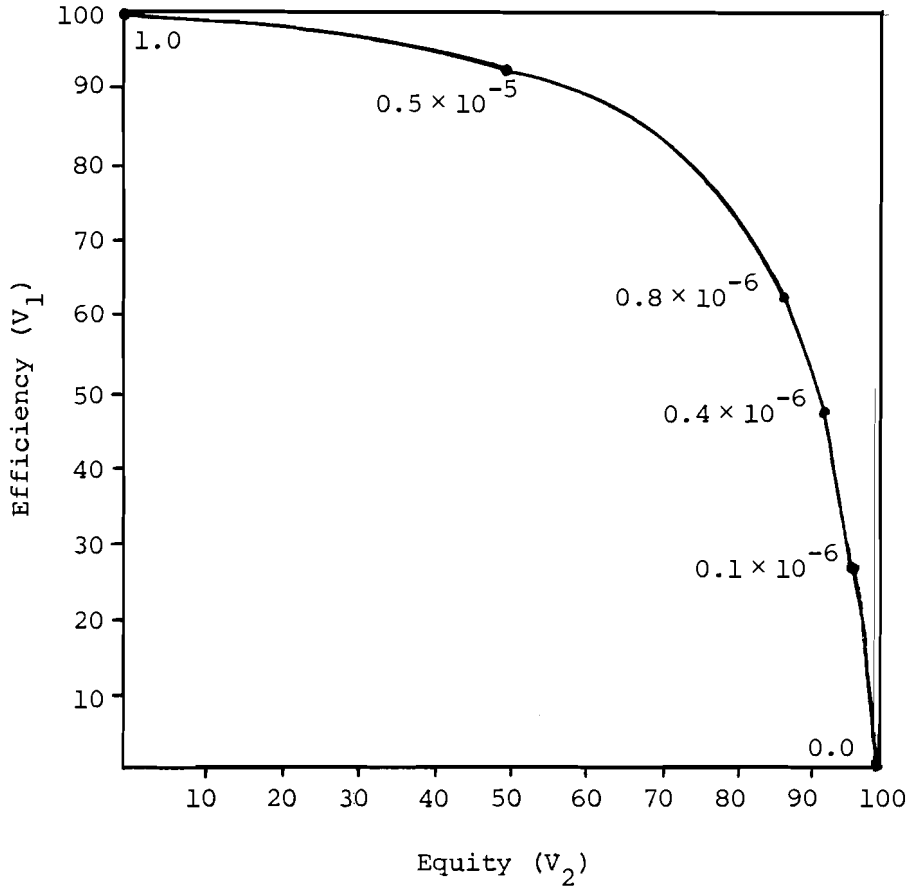

Figure 11. The trade-off curve for efficiency versus equity for different values of the trade-off parameter.

based purely on notions of efficiency or equity (as they have been defined here) and to see how the predicted resource configuration changes with the size of the trade-off parameter.

7. CONCLUSIONS

This paper has considered four criteria of resource allocation in a health care system where size and structure of the population and the availability of resources can change over time and space. These criteria are based on simple notions of patients choice behavior that can be described by a simple attraction constrained gravity model. This model assumes that

there are insufficient resources in the health care system to supply all needs, and that service levels in areas of residence would be strongly influenced by the local availability of resources. The methods are designed with the strategic planning of health care services in mind, in which planners are interested in mainly the broad distributional effects of different spatial resource configurations and not in the detailed pattern of service provision. The criteria considered are based on measures of equity, efficiency, and two types of accessibility, with bounds on the sizes of the facilities allocated in each place of treatment. They have been thoroughly tested on data from the London area of the United Kingdom, which is known to have a very complex distributional problem. As a result of these considerations, accessibility as an operational allocative criterion has been rejected in favor of the equity and efficiency measures. But because it was shown that a regional health care system cannot attain an equitable and efficient allocation of resources simultaneously, it was suggested that the criteria could be merged into a bi-objective trade-off function that allowed the user to test resource configurations trading off one criterion against the other using a trade-off parameter. This was successfully tested on the same data using a purpose designed algorithm based on a modified Frank-Wolfe method. An unsolved problem with this approach, however, was the interpretation of the trade-off parameter since the component objectives were not expressed in compatible units. This aspect needs further work for the multi-objective allocative approach to be completely successful. For more detailed planning purposes, it would also be interesting in the future to develop the methods presented here so that they can apply to multi-level systems, structured in a hierarchical way, that explore equity and efficiency problems when there are multiple services and a range of facility sizes.

APPENDIX: ACCESSIBILITY TAX

The basic model is

$$T_{ij} = D_j B_j W_i e^{-\beta c_{ij}} \tag{A1}$$

The service-need ratio is given by

$$\alpha_i = \sum_j \frac{T_{ij}}{W_i} = \sum_j B_j D_j e^{-\beta c_{ij}} \tag{A2}$$

$$\text{where} \quad B_j = \left[\sum_i W_i e^{-\beta c_{ij}} \right]^{-1} \tag{A3}$$

The equity criterion requires α_i = constant $\forall i$ (i.e., = α)
Define an accessibility tax ρ_i, then

$$\alpha = \sum_j B_j D_j e^{-\beta c_{ij}} \phi_i \tag{A4}$$

where

$$\phi_i = e^{-\rho_i} \tag{A5}$$

and

$$B_j = \left[\sum_i We^{-\beta c_{ij}}\ \phi_i\right]^{-1} \tag{A6}$$

From (4)

$$\phi_i = \alpha / \sum_j B_j D_j e^{-\beta c_{ij}} \tag{A7}$$

In effect, equation (A7) means that zones with a higher accessibility to services will be charged more "tax" than those with lower accessibilities. ϕ_i occurs on both sides of equation (A7), and so it must be found by the iterative sequence

$$\phi_i^{(n+1)} = \alpha / \sum_j B_j D_j e^{-\beta c_{ij}}\ \phi_i^{(n)} \tag{A8}$$

where n is the iteration number. The tax is expressed in the same units as c_{ij}. A problem, however, is to give it an operational meaning. In fact, on closer examination the tax need not be a monetary tax in the traditional sense at all. Non-monetary costs, for example, are incurred by people who are forced to "queue" for treatment on waiting lists. Thus ϕ_i may be used to determine annual patient quotas from different origin zones with the usual provisions giving emergency cases priority. Such a scheme, it may be argued, would distribute the burden of waiting time more fairly among the population as a whole. However, while the idea of a tax is of theoretical interest, there might be political and administrative difficulties associated with its implementation.

REFERENCES

Ben-Akiva, M., and S.R. Lerman (1978) Disaggregated travel and mobility-choice models and measures of accessibility. Pages 654-679, in *Spatial Interaction Theory and Planning Models*, edited by A. Karlqvist, L. Lundqvist, F. Snickars, and J.W. Weibull. Amsterdam: North Holland Publishing Company.

Cochrane, (1975) A possible economic basis for the gravity model. *Journal of Transport Economics and Policy* 9(1): 34-49.

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

Coelho, J.D. (1980) *Optimização, Interacção Espacial e Teoria do Comportamento, Nota no. 5.* Lisbon, Portugal: Centro de Estatistica and Aplicacões, Dept. Matematica Aplicada, Faculdade de Cîencias de Lisoba.

Cohen, M.H. (1961) The relative distribution of households and places of work: a discussion of a paper by J.G. Wardrop. Theory of Traffic Flow. *Proceedings of the Symposium on the Theory of Traffic Flow*, edited by R. Herman. Warren, Michigan, USA. Elsevier.

Dacey, M.P., and A. Norcliffe (1977) A flexible doubly-constrained trip distribution model. *Transportation Research* 11: 203-204.

Domencich, T., and D. McFadden (1975) *Urban Travel Demand: A Behavioral Analysis.* Amsterdam: North Holland Publishing Company.

DHSS (1976) *The NHS Planning System.* Department of Health and Social Security. London: Her Majesty's Stationery Office.

Feldstein, M.S. (1963) Economic Analysis, Operational Research, and the National Health Service. *Oxford Economic Papers* 15:19-31.

Fletcher, R. (1970) *A FORTRAN Subroutine for Quadratic Programming.* United Kingdom Atomic Energy Authority Research Group Report, AERE R6370, UK.

Fletcher, R. (1971) A General Quadratic Programming Algorithm. *Journal of the Institute of Mathematics and Its Applications* 7:76-91.

Frank, E., and P. Wolfe (1956) An algorithm for quadratic programming. *Naval Research Logistics Quarterly* 3:95-110.

HMSO (1944) *A National Health Service.* Cmd. 8502 London: Her Majesty's Stationery Office.

Jefferson, T.R., and C.H. Scott (1979) The Analysis of Entropy Models with Equality and Inequality Constraints. *Transportation Research* 13B:123-132.

LHPC (1979a) *Acute Hospital Services in London.* A profile by the London Health Planning Consortium. London: Her Majesty's Stationery Office.

LHPC (1979b) *The Data Base.* London Health Planning Consortium Study Group on Methodology. Published by the North East Thames Regional Health Authority, UK.

Leonardi, G. (1978) Optimum facility location by accessibility maximizing. *Environment and Planning A* 11:1287-1305.

Leonardi, G. (1980a) *A Unifying Framework for Public Facility Location Problems.* WP-80-79. Laxenburg, Austria: International Institute for Applied Systems Analysis. (Forthcoming in *Environment and Planning A.*)

Leonardi, G. (1980b) *A Multiactivity Location Model with Accessibility- and Congestion-Sensitive Demand.* WP-80-124. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Leonardi, G. (1981) *The Use of Random-utility Theory in Building Location-Allocation Models.* WP-81-28. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Mayhew, L.D. (1979) *The Theory and Practice of Urban Hospital Location*. Ph.D. Thesis Birkbeck College. London: University of London.

Mayhew, L.D. (1980) *The Regional Planning of Health Care Services: RAMOS and RAMOS^{-1}*. WP-80-166. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Mayhew, L.D. (1981) *DRAMOS: A Multi-category Spatial Resource Allocation Model for Health Service Management and Planning*. WP-81-39. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Mayhew, L.D., and A. Taket (1980) *RAMOS: A Model of Health Care Resource Allocation in Space*. WP-80-125. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Mayhew, L.D., and A. Taket (1981) *RAMOS: A Model Validation and Sensitivity Analysis*. WP-81-00. Laxenburg, Austria: International Institute for Applied Systems Analysis. Forthcoming.

McMillan, C. (1975) *Mathematical Programming*. Second edition. New York, N.Y.: John Wiley.

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.

RAWP (1976) *Sharing Resources for Health in England*. Report of the Resource Allocation Working Party. London: Her Majesty's Stationery Office.

Toregas, C., R. Swain, C. ReVelle, and L. Bergman (1971) The location of emergency service facilities. *Operations Research* 19(5):1363.

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. (1967) A statistical theory of spatial distribution models. *Transportation Research* 1:253-269.

Wilson, A.G., and R.M. Kirwan (1969) *Measures of Benefit in the Evaluation of Urban Transport Improvements*. WP 43. London: Centre for Environmental Studies.

Wilson, A.G. (1971) A family of spatial interaction problems and associated developments. *Environment and Planning* 3:1-32.

Wilson, A.G. (1974) *Urban and Regional Models in Geography and Planning*. London: Wiley.

RECENT PUBLICATIONS IN THE HEALTH CARE SYSTEMS TASK

Hughes, D.J. (1979) *A Model of the Equilibrium between Different Levels of Treatment in the Health Care Systems: Pilot Version* (WP-79-15).

Fleissner, P. (1979) *Chronic Illnesses and Socio-Economic Conditions: The Finland Case 1964 and 1968* (WP-79-29).

Shigan, E.N., D.J. Hughes, and P. Kitsul (1979) *Health Care Systems Modeling at IIASA: A Status Report* (SR-79-4).

Rutten, F.F.H. (1979) *Physician Behaviour: The Key to Modeling Health Care Systems for Government Planning* (WP-79-60).

A Committee Report (1979) to IIASA by the participants in an Informal Meeting on *Health Delivery Systems in Developing Countries* (CP-79-10).

Shigan, E.N., P. Aspden, and P. Kitsul (1979) *Modeling Health Care Systems: June 1979 Workshop Proceedings* (CP-79-15).

Hughes, D.J., E. Nurminski, and G. Royston (1979) *Nondifferentiable Optimization Promotes Health Care* (WP-79-90).

Rousseau, J.M., R.J. Gibbs (1980) *A Model to Assist Planning the Provision of Hospital Services* (CP-80-3).

Fleissner, P., K. Fuchs-Kittowski, and D.J. Hughes (1980) *A Simple Sick-leave Model used for International Comparison* (WP-80-42).

Aspden, P., R. Gibbs, and T. Bowen (1980) *DRAM Balances Care* (WP-80-43).

Aspden, P., and M. Rusnak (1980) *The IIASA Health Care Resource Allocation Submodel: Model Calibration for Data from Czechoslovakia* (WP-80-53).

Kitsul, P. (1980) *A Dynamic Approach to the Estimation of Morbidity* (WP-80-71).

Shigan, E.N., and P. Kitsul (1980) *Alternative Approaches to Modeling Health Care Demand and Supply* (WP-80-80).

Hughes, D.J., and A. Wierzbicki (1980) *DRAM: A Model of Health Care Resource Allocation* (RR-80-115).

Aspden, P. (1980) *The IIASA Health Care Resources Allocation Submodel: DRAM Calibration for Data from the South West Health Region, UK* (WP-80-115).

Mayhew, L., and A. Taket (1980) *RAMOS: A Model of Health Care Resource Allocation in Space* (WP-80-125).

Mayhew, L.D. (1980) *The Regional Planning of Health Care Services: RAMOS and RAMOS^{-1}* (WP-80-166).

Pauly, M.V. (1981) *Adding Demand, Incentives, Disequilibrium, and Disaggregation to Health Care Models* (WP-81-4).

Mayhew, L.D. (1981) *DRAMOS: A Multi-category Spatial Resource Allocation Model for Health Service Management and Planning.* (WP-81-39).