Evaluating and Improving the Accessibility of Primary Health Care Services

by

Robert L Bowerman

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 1997

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

Investment in human resources is one of the key factors underlying the development of society. With human resource investment it is possible both to strengthen the development of a society and to increase its social equity. Primary health care is a critical investment in human capital that can bring about progressive societal change. Several factors, however, affect the provision of these health services in developing countries, including limited public funds available for providing services and a need for these services to target a population that is both poor and geographically dispersed. Since primary health care is an essential service, it is important to evaluate its provision to determine whether the service satisfies the needs of the target population and is distributed justly among the intended users.

This thesis examines the problem of evaluating and improving the potential accessibility of a target population to primary health care services. Towards this end, it develops a generic model of potential accessibility. It also examines how spatial aggregation of the target population can lead to errors in the evaluation of accessibility, and discusses methods of disaggregating population counts to a grid to reduce this spatial aggregation error.

Further, it develops a generic Accessibility Optimization Problem (AOP) that takes a facility-oriented approach to improving accessibility. Two subproblem formulations are also discussed for the AOP. The Facility Location Subproblem (FLS) adjusts the facility configuration to improve the efficiency and equity in the distribution of accessibility among the target population while the Resource Allocation Subproblem (RAS) modifies the allocation of resources to existing facilities. Specific accessibility optimization models for the minimum distance accessibility measure and the Joseph and Bantock [1982] accessibility measure are developed from the generic formulations. These accessibility measures are used to evaluate the current accessibility, and the optimization models are applied in two specific planning scenarios to examine potential strategies of improving accessibility to family planning services in the Central Valley of Costa Rica.

Acknowledgements

First, I wish to thank my thesis advisers, Professor Paul H. Calamai and Professor G. Brent Hall, for their sage advice in all aspects of this thesis. Their expert guidance greatly strengthened this thesis in all matters from the detailed technical aspects to matters of style and presentation. Their encouragement, support, and guidance were instrumental in the completion of this thesis. I would like to take this opportunity to thank the members of my comprehensive examination committee, Professors M. Chandrashekar, P. Kanaroglou, and K. Ponnambalam for the critical advice they gave me on my thesis proposal. In addition, I wish to thank Professors O. Basir, B. Boots, and A. Joseph for their thoughtful questions and comments made during the thesis defence and to express my appreciation to Professor C. Macgregor for acting as a delegate for K. Ponnambalam during the defence.

This research would not have been possible without the support of the International Development Research Centre of Canada (Grant 92-1152-01 "Redatam+GIS Generic Population-Related Application Tools") and the Natural Sciences and Engineering Research Council (Research Grant OGP0005672 and Equipment Grants EQP0107889 and EQP0173530). I am also indebted to the Natural Sciences and Engineering Research Council and the University of Waterloo for the scholarships that these institutions provided during the course of this thesis. I would also like to acknowledge Professors L. Rosero and V. Gomez of the University of Costa Rica for their assistance in providing the information used for the empirical example in this thesis. Further, I wish to thank Dr. Arthur Conning and Serge Poulard of the Latin American Demographic Centre in Santiago, Chile for their cooperation during the course of this research.

I wish to thank a number of friends and colleagues, who assisted and encouraged me in various ways. Both Robert Feick and Gunnar Hillgartner worked on the Redatam+GIS project with me and provided continuing encouragement to complete this research. Further, the long hours spent in the lab were made easier with the good spirits of Dr. John Hodgson, Layi Oshinowo, Aaron Baumal, and Henry Venema. I also wish to thank Annette Dietrich in taking care of many of the formalities associated with being a doctoral candidate.

I am indebted to my family for their understanding and patience. They were a great sustaining influence, both morally and financially, throughout my academic career at the University of Waterloo. And finally, to my wife, Christine, I give my warmest thanks and dedicate this thesis to her. Her tolerance and unconditional support were invaluable and she gave me the drive necessary to complete this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Primary Health Care Provision in the Developing World

Investment in human resources is one of the key factors underlying the development of a society. With human resource investment it is possible to strengthen both the development of a society and to increase its social equity [ECLAC, 1992]. Primary health care is a critical investment in human resource development that can bring about progressive social change. Moreover, it constitutes one of the most important welfare functions of government.

In developed countries, the health care system is generally characterized as a three-level hierarchical system [Joseph and Phillips, 1984]. The lowest level of this hierarchy is primary health care which provides the first point of contact between patients and the health care system. Providers of primary health care include general practitioners, public health clinics, nurses, health auxiliaries and hospital emergency rooms. The secondary level consists of more specialized care provided by general hospitals and specialist physicians. The highest, or tertiary, level consists of institutions providing highly specialized care such as specialist hospitals and clinics. Fendall [1981] proposes a similar health delivery system for developing countries consisting of five layers: village centres and rural health clinics for primary care, district and regional centres for the second level, and a national health centre providing tertiary care.

The World Health Organization (WHO) considers primary health care to involve a form of service delivery based on equity, intersectoral action, and community participation for the provision of essential health care [Tarimo, 1991]. In the same context, the

International Conference on Primary Health Care held at Alma Ata in 1978 [WHO and UNICEF, 1978] characterized essential care as consisting of at least: the treatment of common diseases and injuries; maternal/child care and family planning; the provision of essential drugs; immunization; the control of communicable diseases; health education; an adequate supply of safe water and basic sanitation; and, an adequate food supply and proper nutrition. These components and the concept of universally accessible essential health care are considered by WHO to be the key for achieving their goal of health for all by the year 2000 [WHO, 1981].

A critical component of essential health care, as defined above, is maternal-child health and family planning. A woman's control of her fertility can be considered the "freedom from which other freedoms flow" [WHO, 1992, p. 4]. Without this control, it is difficult for a woman to complete her education and stay employed while making independent marital decisions. In addition, a woman's fertility control has strong health benefits in reducing high-risk and unwanted pregnancies. The United Nations' Economic Commission for Latin America and the Caribbean (ECLAC) considers that investments in maternal-child health and family planning resources are critical for increasing both the development and social equity of a society [ECLAC, 1992]. Consider the following example given by ECLAC. Many women, particularly among disadvantaged segments of the population, have a higher fertility rate than they desire and have more children to raise with insufficient economic support. These same disadvantaged women often have poor access to medical care during pregnancy and delivery. This poor accessibility leads to infant malnutrition and higher rates of mortality in both infants and mothers. Poor infant nutrition, combined with poor access to educational opportunities, significantly limits the intellectual and physical development of these children and, hence, their future. These children enter the labour market at an early age in casual jobs of low productivity and, perpetuate this vicious cycle of inequity. Therefore, investments in maternal-child health and family planning programs help to reduce the intergenerational cycle which perpetuates economic marginality and social exclusion [ECLAC, 1992].

In spite of its great importance in the sustainable development of a society, the provision of primary health care in developing countries is often beset by numerous problems. Public funds available for providing services are limited; in many countries, government expenditure on health programs is less than 2% of the gross national product [Tarimo, 1991]. Further, the need for these services is unevenly distributed over

space, making the goal of universal accessibility difficult to achieve. Since primary health care is an essential service, it is important to evaluate its provision to determine whether target populations are adequately covered, whether services satisfy the needs of the target population, and if they are distributed justly among the intended users.

There are important barriers to achieving effective planning for primary health care. In many developing countries, health care planning has tended to be episodic and capricious so that local changes in the health care system, are often implemented without consideration of the effects of the changes on the whole. Health care systems are often planned in a fragmented manner through funding provided by public, private, charitable, and aid sources. This problem is compounded by the fact that not only are resources inadequate but they are sub-optimally located with respect to demand [Oppong and Hodgson, 1994]. An additional difficulty relates to the complexity of the planning process. As mentioned above, it is important to ensure that health care is accessible to a widely dispersed and unevenly distributed population. The evaluation of health care accessibility is a complicated problem and providing accessible health care in an equitable manner is a complex issue with both spatial and aspatial aspects. Phillips [1990] notes that difficulties in improving health care services are not solely related to the scarcity of financial resources but also result from practical difficulties in management and planning.

1.2 Evaluating and Improving the Accessibility of Primary Health Care

This thesis is concerned with the problem of access to primary health care, with particular emphasis on, but not limited to, developing countries. Mathematical models and techniques are developed to assist health care planners and decision makers to evaluate and improve the accessibility to primary health care services.

Evaluating accessibility is important for assessing the current state of the health care system. Determining the current patterns of accessibility allows decision makers to identify areas and regions that have a deficient supply of or access to primary health care. Further, the concept of accessibility evaluation can be broadened to consider not only the spatial distribution of accessibility but also the differential accessibility to primary health care services among different target groups. Thus, accessibility evaluation

allows planners and decision makers to visualize the strengths and weaknesses of the current configuration of the primary health care system relative to the spatial distribution of demand for services.

Being able to diagnose accurately the nature of problems, with respect to accessibility, in a study area is only a partial solution. Beyond assessing existing accessibility, decision makers require assistance in developing strategies to improve the efficiency and equity of primary health care provision. In fact, the World Health Organization (WHO) [1994] recognizes that:

> Population-based socioeconomic, cultural, demographic and epidemiological information is vital for choosing priority areas for action, planning public health interventions and evaluating progress. However, to improve the implementation of services and programmes, better service-based information is required at the district level. To improve health status and achieve greater equity, district health services and public health programmes need to be efficient, have high coverage and be of good quality. Only when these three requirements are fulfilled will the full potential of public health action be realized (p. 2).

Thus, models and information systems that can evaluate the accessibility of primary health care services and assist local decision makers in planning new services provide an important link in improving the overall level of public health and development in a region.

Present in virtually all analyses of access to primary health care are the concepts of the efficiency of a health care system and the equity in accessibility for all potential users. While there is general agreement as to the conceptual and operational definitions of these terms, they can be assessed in a number of different ways. Conceptually, efficiency can be thought of, in the context of this thesis, as measuring the total aggregate level of accessibility, or benefits derived from accessibility, of the target population to the health care system. Equivalently, equity can be thought of as examining the fairness, impartiality, or equality of the service provision. Thus, a health care system would be considered more efficient if resources were allocated to areas where they have the maximum aggregate benefit, whereas in a more equitable system there would be reduced variation in access between areas and population subgroups. These concepts, as well as specific methods of assessing them, are further discussed in Chapter 5.

There are many approaches available for improving a population's access to primary health care. The emphasis in this thesis is on a service- or facility-oriented approach to improving spatial accessibility to services. Using this approach, there are two main methods available to improve or equalize access to services: either by reducing the distance deterrence between supply points and population or by increasing the resources available at supply points. These two methods correspond to two distinct modelling approaches. First, access can be improved through determining the locations of new facilities that are optimal in terms of one or more criteria. This is a facility-location or location-allocation approach and there are several examples of the use of this approach for primary health care planning in the developing world. The use of this approach has been quite contentious and has been criticized by some authors such as Rondinelli [1985] and Gore [1991a; 1991b]. These criticisms and the appropriateness of this approach are discussed further in Chapter 5. The second approach is to find the best resource allocations among existing facilities. There is no published evidence of this approach having been applied to primary health care planning in the developing world.

It should be emphasized that the facility-oriented approach is not the only approach available for improving accessivilirt to services. Other issues such as poor education, housing, and sanitation may be crucial barriers to achieving adequate access to primary health care [Tarimo, 1991]. Furthermore, the low quality of services provided at existing service providers may also be an important barrier to access of primary health care services (see, for example, Annis [1981]). However, these types of problems are hard to quantify, vary dramatically in importance in different areas, and generally require a high degree of knowledge of local conditions. Thus, they are not amenable to being formulated within a generic mathematical framework. On the other hand, the models formulated in this thesis are more generic and less dependent on local conditions.

Although the strategies for improving accessibility developed by these models are not necessarily appropriate for every situation, they provide very important information to planners and decision makers in cases where additional resources are being allocated to a health care system or new facility locations are being planned. In these situations, it is important to allocate resources optimally and to locate new service providers so that scarce resources are used most effectively when a system is expanded [Phillips, 1990]. Furthermore, these models can also be used to generate an optimal configuration of resource allocations and facility locations and to compare this optimal configuration

to the current system. The information provided by this particular analysis is a very important tool in examining the effectiveness of the current system configuration.

1.3 Objectives of this Thesis

The specific objectives of this thesis are as follows.

1. To describe, and specify formally, a generic model of potential accessibility to primary health care services in developing countries. This generic model can incorporate both spatial influences on accessibility and aspatial influences such as socio-demographic or health system organizational variables.
2. To discuss how spatial aggregation of the target population for primary health care services can affect the evaluation and improvement of accessibility and to discuss methods for reducing the effects of aggregation error.
3. To develop, specify, and apply optimization models and solution techniques to assist decision makers in improving efficiency and equity in the allocation of resources within a system providing primary health care services.

1.4 Outline of this Thesis

In order to accomplish the objectives described above, this thesis is organized into eight chapters. The outline below summarizes the contents of the remaining seven chapters.

Chapter 2 provides a review of existing literature on accessibility to primary health care emphasizing potential spatial accessibility measures. As well, two models for improving accessibility are described briefly: facility location models, and spatial interaction models. The application of these models to health care planning, particularly for primary health care in developing countries, are also described.

Chapter 3 specifies a formal generic model of accessibility to primary health care services for an individual. A behavioural interpretation of this model is then presented in the context of individual choice theory.

Chapter 4 considers the effects of spatial aggregation of demand. Although the model of accessibility specified in Chapter 3 defines accessibility in terms of the individual consumer of health services, the operationalization of this model on a regional scale to both evaluate and improve accessibility requires the spatial aggregation of these individuals. This chapter examines the issues and difficulties related to spatial aggregation.

Chapter 5 discusses the appropriateness of using optimization models to improve primary health care accessibility. A generic accessibility optimization model is derived and potential planning objectives are specified for examining both the equity and the efficiency of the health care delivery system.

Chapter 6 develops three specific models and associated objectives based on the generic formulation provided in Chapter 5 for two measures of potential spatial accessibility.

Chapter 7 applies the concepts and models developed in the thesis to a data set on family planning consultations for women in the fertile age cohort (15–49 years) living in the Central Valley of Costa Rica.

Chapter 8 concludes the research, summarizes the contributions of the thesis, discusses the results presented in the previous chapter, and examines potential directions for future research.

Chapter 2

Existing Accessibility Measures and Improvement Methods

This chapter reviews health care accessibility measures, spatial interaction models, and facility location models. These three topics are central to the development of a framework for evaluating and improving accessibility to primary health care services in developing countries.

The first section describes pertinent concepts for evaluating health care accessibility. Khan and Bhardwaj's [1994] conceptual framework of access to health care is used to organize the discussion, which focuses on studies that emphasize geographical accessibility to primary health care in developing countries. Section 2.2 and 2.3 review two different families of models that are used in developing methods and strategies for improving accessibility. Section 2.2 deals with spatial interaction models, which quantify the level of interaction or flows between an area or group of origins to various destinations in a system. These models have an obvious application to the planning of health care systems as they can be used to estimate patient flows. Several examples of spatial interaction models applied to health care planning in *developed* countries are discussed as there are no known examples of their application to health care planning in *developing* countries.

Section 2.3 examines facility location models. As mentioned in the previous chapter, facility location models seek the optimal locations for service distribution with respect to one or more criteria, the current distribution of demand, and the current system configuration. This section surveys several existing facility location models and their

solution techniques. Included in this discussion are both traditional models, with their assumption of nearest-centre allocation, and more recently developed models that are based on combining facility location models and spatial interaction models. Examples are then given of applying facility location models to primary health care planning in developing countries.

2.1 Accessibility to Health Care

Health is a vital factor in determining the productivity, development, and well-being of a society. Therefore, provision of health care is one of the most important influences on a society and, for publicly-provided health services, one of the most important welfare functions of government. The distribution and utilization of resources in the health care system have a great impact on society. This is particularly true for the disadvantaged members of society where poor personal health plays an important role in poverty and deprivation [Knox, 1979]. It is therefore highly desirable that health care services be available to all members of society.

The demand for health care services, like the demand for most public services, comes from spatially-dispersed individuals so that demand is unevenly distributed over space [Dear, 1974]. Public health care services, however, are distributed from discrete service facilities with fixed locations. Therefore, completely equal availability of the health care system could only occur if every individual had immediate and uninterrupted access to a health care facility [Joseph and Phillips, 1984].

The variation in the distance of individuals to the nearest health care facility causes differential accessibility to health care services. Furthermore, additional factors, which vary among individuals and instituitions, such as psychological, socio-economic, cultural, and organizational factors also affect accessibility. Thus, it is not possible to have a completely equal health care system; it is, however, important to have an equitable one. This would be a health system that is easily accessible to the target population.

2.1.1 A Framework for Accessibility

Health care accessibility can embody multiple dimensions and be influenced by many factors. The general question of what is accessibility, particularly with respect to health care, has been the subject of much debate. Phillips [1990] notes that the notion of ac-

cessibility is a "slippery" concept to define. Nevertheless, there is a large commonality between existing definitions of health care accessibility.

Donabedian [1973] distinguishes two basic forms of accessibility to health care services, geographic and social accessibility. Geographic, or physical accessibility, emphasizes the importance of space or distance as a barrier to access to the health care system. Social or socio-organizational accessibility addresses a variety of issues, separate from but related to, geographic accessibility. It is influenced by social, economic, demographic, and health system organization variables. Several authors, such as Aday and Anderson [1974], have recognized the need to differentiate between the potential availability of services and the actual entry of consumers into the system. Similarly, Joseph and Phillips [1984] made the distinction between potential accessibility, which is measured by spatial and socio-economic aspects of a health care system and realized accessibility or utilization, which is the actual use of the system.

Khan and Bhardwaj [1994] place the geographic/social dimensions and the potential/realized dimensions of accessibility into a conceptual framework of access to health care and they propose a typology of access, as illustrated in Figure 2.1. In their model, accessibility is moderated (negatively) by barriers and (positively) through facilitators that reflect characteristics of both the potential users and the health care system itself. Utilization is greatly influenced by the availability of services as well as the characteristics of both the users and the health care system.

	Spatial	**Social**
Potential	I Potential Spatial Accessibility	II Potential Social Accessibility
Realized	III Realized Spatial Accessibility	IV Realized Social Accessibility

Figure 2.1: A typology of access after Khan and Bhardwaj [1994]

Khan and Bhardwaj [1994] also distinguish between geographic (or spatial) accessibility and social (or aspatial) accessibility. Geographic accessibility considers how space

affects the availability of health services, while, consistent with Donabedian [1973], social accessibility relates the availability of services to non-geographic dimensions of the consumers or the system. It is important to note that both geographic and social accessibility can have both spatial and aspatial patterns. Thus, the spatial distribution of the relative availability or use of health care services, whether defined in terms of geographic or social accessibility, is reflected in a spatial pattern of accessibility. Aspatial patterns are manifested in the differential availability or use of health care services among various sub-groups in a population due to economic, social, cultural, religious, political, psychological, and other barriers. Within this model, Khan and Bhardwaj [1994] differentiate four types of accessibility, potential geographic, potential social, realized geographic, and realized social accessibility.

The next four subsections survey existing models of health care accessibility within these four categories. Where appropriate, formulae for the accessibility measures are given. Studies examining accessibility to maternal-child health/family planning services are then discussed.

2.1.2 Potential Geographic Accessibility

Measures of potential geographic accessibility concentrate on the spatial configuration of service providers relative to the spatial distribution of relevant population groups. Several methods have been developed for measuring this form of access to primary health care services. These methods have been used in a variety of contexts, including intra-urban and regional areas, in both developed and developing countries. For these measures, the emphasis has been on population groups defined by geographic regions or sub-regions.

One of the simplest measures of potential geographic accessibility, the minimum distance accessibility measure is defined as the distance of potential consumers to the nearest facility. Mathematically, the accessibility of population group i can be expressed as

$$A_i = \min_j D_{ij}, \tag{2.1}$$

where A_i is the accessibility of population group i, and D_{ij} is the distance between population group i and service provider j. Okafor [1990] uses this measure to quantify the

geographical accessibility of population groups to general hospitals in rural Nigeria. Although Okafor acknowledges this to be a crude measure, difficulties with the acquisition of quality data prevented the use of a more appropriate measure. Annis [1981] uses a similar approach and defines geographic accessibility in terms of distance to the nearest centre for a region of rural Guatemala, although his study concludes that quality of service was more important in this case than distance to the facility.

A more sophisticated approach to measuring physical potential accessibility of health care services under the constraints of data collection and data quality problems was used in Nigeria by Ayeni *et al.* [1987]. Current accessibility of the health care system is approximated by allocating each settlement to the nearest health care facility and calculating the average straight line distance to this facility. Then the average distance to the nearest facility is recalculated after facilities are optimally located using a p-median facility location model (see section 2.3.1). These average distance values are then converted into utilization estimates using a distance decay function and used to estimate the loss in utilization due to spatially inefficient placement of the health care facilities. In their case study, these authors estimated that there was a 23% loss in utilization of maternal-child health centres and a 25% loss in dispensaries in 1979. Oppong [1992] follows the same approach for a district of Ghana but extends it using both a p-median model as well as population covering models and a three-level service hierarchy for the different types of facilities.

The previous accessibility measures are predicated on the assumption that people attend the nearest facility. Often people do not attend the nearest facility and instead may choose a different facility (see, for example, Martin and Williams [1992] or Bailey and Phillips [1990]). A different approach for measuring potential geographic accessibility, the gravity model, relaxes the nearest-centre assumption. This approach incorporates a mathematical function, termed a distance decay function, to model the frictional effect of distance. There are two main families of distance decay functions [Fotheringham and O'Kelly, 1989]: exponential functions, $f_D(\beta, D) = \exp(-\beta D)$, and power functions, $f_D(\beta, D) = D^{-\beta}$, where D is the distance and β is the decay parameter. The formulations presented in this chapter utilize the distance decay functions used in the original research. It should be noted, however, that these measures can be easily re-formulated in terms of alternative distance decay functions. This is discussed further in the context of spatial interaction models.

For the gravity model approach, the accessibility of a particular population group is

calculated as the sum of the resources available at the different service centres weighted by the distance decay function. Martin and Williams [1992] give a good example of the gravity model approach in examining the accessibility of primary health care provided by general practitioners in the City of Bristol in the United Kingdom. The population groups in their study are defined by 100 metre grid cells. Using this information, they calculate accessibility surfaces to general practitioners using the classic Hansen [1959] accessibility measure (as well as several other related measures)

$$A_i = \sum_j S_j \exp\left(-\beta D_{ij}\right), \tag{2.2}$$

where S_j is the size of, or resource level at, service provider j. In this study, S_j is equal to the number of general practitioners at a given surgery and D_{ij} is taken to be the straight-line distance between the centre of grid cell i and surgery j. Furthermore, the distance decay function, calibrated using a spatial interaction model, is defined as $f_{ij}^D = \exp\left(-1.57 D_{ij}\right)$. For their study area, they find very few areas with poor accessibility to primary health care services.

Knox [1978; 1979] refines the basic gravity model approach by using levels of car ownership as a surrogate for the relative mobility of the population and applies this measure at the neighbourhood level in Scottish cities adjusting the accessibility measure by the average travel speeds for cars and public transit. The new travel-time accessibility estimate is calculated by

$$A_i = \left(\frac{X_i^C}{T^C} + \frac{X_i^{NC}}{T^{NC}}\right) \sum_j S_j \exp\left(-\beta D_{ij}\right), \tag{2.3}$$

where X_i^C is the proportion of car-owning households in neighbourhood i, $X_i^{NC} = 1 - X_i^C$ is the proportion of households that do not own a car in a neighbourhood i, T^C is the average time to travel a given distance by car, and T^{NC} is the time to travel the distance by public transport. He finds that variations in the accessibility of primary health care facilities reinforce patterns of social deprivation and medical need.

Joseph and Bantock [1982] propose a modification to the gravity model. Their measure considers the potential demand on service providers and adjusts the resource availability at each facility by this potential demand. Using this measure, accessibility is

defined as

$$A_i = \sum_{j, D_{ij} \leq R_i} (S_j / C_j) \, D_{ij}^{-\beta}, \tag{2.4}$$

where C_j is the potential demand on service provider j and is defined as

$$C_j = \sum_{i, D_{ij} \leq R_i} P_i D_{ij}^{-\beta}. \tag{2.5}$$

Note that this accessibility measure defines a maximum service range, R_i, for each population group. A service provider beyond this range is considered inaccessible and does not affect the accessibility of group i. In their case study using data from Wellington County, Ontario, Joseph and Bantock [1982] use a distance decay exponent, β, of 2 based upon empirical work by others, and the accessibility index is calculated for R_i set to 5.0 miles, 10.0 miles, and 15.0 miles. They find that the measure is not very sensitive to changes in the service range once all areas are in the range of a general practitioner. As well, the study shows greater potential accessibility to general practitioner services in areas near urban centres, although this is mitigated somewhat by physicians in rural areas having fewer potential clients.

Other authors have also used this approach. For example, Rosero-Bixby [1993] follows the Joseph and Bantock approach in evaluating the potential geographic accessibility to health care facilities in Costa Rica. In addition, a slightly modified version of this approach is used by Khan [1992] and applied in Ohio.

2.1.3 Potential Social Accessibility

In contrast to potential geographic accessibility, measures of potential social accessibility examine the differential availability of health care resources emphasizing the importance of socio-demographic and organizational factors. These measures examine the possible usage of primary health care services, rather than actual utilization behaviour. As opposed to the measures proposed in the previous section, potential social accessibility measures do not explicitly consider the effects of distance or travel cost.

Potential social accessibility measures typically involve calculating the ratio of the supply of health care to the demand for health services, often at a regional level. These ratios are termed regional availability measures by Joseph and Phillips [1984]. Calculations of regional health care accessibility assume that the boundaries for a region

are impermeable. This approach, therefore, becomes increasingly difficult to use or can give misleading or inappropriate results when using highly spatially disaggregate data [Joseph and Phillips, 1984]. Many studies attempt to use these measures to quantify equity in resource allocation between rural and urban areas or between different ethnic groups in a multi-racial country [Akhtar and Izhar, 1986; Khan, 1985; Okafor, 1987].

One advantage of ignoring distance with potential social accessibility is that these methods require significantly less data than those that consider potential geographic accessibility. Hence, they are widely used in developing countries with data deficiencies since they are very useful in determining whether resources are distributed equitably between regions within a country.

As mentioned previously, the basic potential measure of social accessibility is a ratio of supply to demand, namely

$$A_i = S_i / P_i, \tag{2.6}$$

where S_i is the total level of resources available in region i and P_i is the total target population or some other surrogate for the "need" for health care in region i. Note that S_i is some measure of the total level of the resource of interest – such as doctors, consultation hours, number of health posts – located within a given region.

Examples of the use of these ratios are widespread. In India, Akhtar and Izhar [1986] examine variations in hospital facilities and hospital beds per capita at the district level. They discovered severe regional imbalances with a ten-fold difference between the best and the worst districts. In Sierra Leone, Stevenson [1987] calculates ratios of hospital beds, doctors, and nurses per capita and also finds that health services were very unevenly distributed. Mesa-Lago [1985] also finds a wide disparity in the ratio of physician, hospital beds, and medical visits per capita between the provinces in Costa Rica even though Costa Rica is "one of the few countries in Latin America with almost universal coverage in health care" [Mesa-Lago, 1985].

Khan [1985] introduces several potential social accessibility indices and calculates them for 62 subdivisions in Bangladesh for urban, rural, and combined populations. The indices relate to the proportion of the population which has access to outpatient services, the estimated quantity of hospital beds utilized per unit of population, the proportion of facilities offering satisfactory services, and a composite index of relative

access to health care. Khan finds that there is both a significant variation in these indices among the subdivisions and a large urban/rural disparity, with rural areas being significantly worse.

Joseph and Phillips [1984] illustrate the use of a location quotient as an accessibility measure. It is defined as follows:

$$A_i = \frac{S_i/P_i}{\sum_i S_i/\sum_i P_i}. \tag{2.7}$$

A location quotient measures a region's proportion of resources relative to its proportion of the target population. If this accessibility measure is greater than 1.0 then that region is supplied with more resources than the average while an accessibility level less than 1.0 indicates an undersupply. Stimson [1980] uses this measure in Australia for examining the concentration of GP services in Adelaide.

It is also useful to calculate system-wide measures of inequity in the distribution of resources or accessibility. One such measure is the coefficient of localization which measures the concentration across regions of an activity or resource of interest relative to the base magnitude [Joseph, 1982]. Again, in the context of measuring the health care accessibility, this coefficient of localization is defined [Joseph and Phillips, 1984] as:

$$Z_{CL} = \frac{1}{2}\sum_i \left| \frac{S_i}{\sum_i S_i} - \frac{P_i}{\sum_i P_i} \right|. \tag{2.8}$$

In interpreting the coefficient of localization, a value of 0.0 indicates that the general practitioners are distributed in the same proportion as the population. Increasing values of the coefficient indicate greater levels of localization. The theoretical upper limit of 1.0 would indicate that the regions containing general practitioners and the regions containing the population are disjoint. For example, Okafor [1987] calculates a coefficient of localization for different types of health facilities in Nigeria. In this study, regional disparities were found with hospitals, health centres, and maternal centres exhibiting a higher level of inequality than dispensaries. He concludes by stressing the importance of the spatial component of health care provision. However, as Joseph [1982] points out, the coefficient of localization must be interpreted as a measure of relative and not absolute concentration.

2.1.4 Realized Geographic Accessibility

While there must be services available in order for a health system to be considered accessible, other factors intervene to affect the overall accessibility of the system [Khan and Bhardwaj, 1994]. Several authors discuss accessibility in terms of utilization patterns. For example, Donabedian measures access to health care by the level of usage in relation to 'need' [1973, p. 211]. Aday and Anderson [1974] define accessibility to be whether people in need of medical service receive it. Thus, as Joseph and Phillips [1984] argue, the utilization of health services and resources by individual consumers reveals the actual accessibility of the system. However, a problem with this approach is the difficulty in defining the need for health care services [Joseph and Phillips, 1984]. A second problem is that, even if the need is known, the examination of utilization makes it difficult to separate the effects of various factors, such as age, gender, and socio-economic status, affecting accessibility since these factors have complex inter-relationships.

Indicators of realized accessibility are generally obtained from population census data or surveys of users and are of two main types: utilization rates and statistical models. Utilization rates are computed by grouping the population by various socio-demographic and spatial factors and then calculating the utilization rate for each group. On the other hand, statistical models, such as the logit model, attempt to quantify the effects that various factors have on health-seeking behaviour. In contrast to the sharp distinction between potential geographical and social accessibility measures, the difference between realized geographic and social accessibility is somewhat blurred as many studies quantify both spatial and aspatial influences on utilization behaviour. However, for the purposes of this discussion, realized geographic measures are those measures that emphasize the importance of the spatial factors on utilization behaviour.

Girt [1973] undertook a study that examines the impacts of distance on usage of cottage hospitals in rural Newfoundland. His study involved surveying 1400 individuals as to whether they would seek medical attention for different conditions. Curves modelling the probability to consult versus distance were then fitted to the survey responses. The results indicate that, in general, the likelihood of using a health care facility decreased with increasing distance to the medical facility. The relationship, however, was not straightforward and distance seemed to have both a positive and a negative effect on the individual consultation rate. Individuals were more sensitive to the development of illness the further away they were from a physician but those furthest

away were less likely to consult due to the additional effort required. To a point, the probability of consultation for some diseases increased with distance from a medical facility. It was hypothesized that the anticipated difficulty, due to distance, of receiving emergency care for these diseases leads to the increasing likelihood of non-emergency consultations. Past this point, the deterrent effect of distance became paramount and the consultation probability decreased with increasing distance.

Recent studies of the effect of distance in primary health care utilization have taken a more sophisticated approach by disaggregating respondents using demographic variables such as age, gender, social status, and mobility. Haynes and Bentham's [1982] study compares utilization rates of general practitioners by adults in East Anglia, England. This study found that, in general, residents in remote regions had a lower consultation rate than those in more accessible regions. Also, the differential utilization rate due to distance decay effects varied among the different demographic categories.

In the developing world, Bailey and Phillips [1990] examine the spatial patterns of primary health care utilization in the metropolitan area of Kingston, Jamaica. In this study, three pairs of sites in close proximity but with contrasting socio-economic composition were selected, with one of each pair being high status and the other low status. Fifty respondents at each site were interviewed as to the mode of transport used, the travel time taken, and the type and proximity of their primary health care provider. The study found that most respondents did not attend the nearest facility and that the attendance patterns varied between the high status sites and the low status sites, even if they were in close proximity. Respondents from low status sites made greater use of public facilities and casualty departments and residents of poor peripheral sites experienced much longer travel times to reach primary health care. Finally, Bailey and Phillips [1990] note that people often do not attend the nearest facility and that "it is important for health service planners to recognize [this]" (p. 11). Thus, this study tends to support the use of gravity model measures for assessing potential geographic accessibility.

2.1.5 Realized Social Accessibility

The final type of accessibility defined by Khan and Bhardwaj [1994] is realized social accessibility. With this type of accessibility, emphasis is on the actual use of health care services and the influence of non-spatial factors such as social, economic, demographic,

and organizational variables, although many of these studies do incorporate the effect of distance on utilization.

Several demographic and social variables, such as age and gender, have been found to have an influence on health care utilization. In general, females and older people have a higher rate of utilization than males and younger people although very young children and young mothers are often frequent users [Phillips, 1986]. However, this relationship depends upon the type of service being offered [Hall, 1988]. Other variables that affect utilization include race, income, mobility, and social class, as well as a variety of factors that affect the individual such as past experience [Bailey and Phillips, 1990].

Studies often examine the utilization of health care services in terms of both demographic variables and physical variables. Examples of this approach include Hall [1988] who uses a multivariate linear regression model in examining the utilization of community mental health centres in Auckland, New Zealand. For this study, the utilization rate per 1000 total population was the dependent variable and socio-demographic variables involving marital status, gender, income, age, and a distance variable, the mean arterial distance travelled, were the independent variables. This model was calibrated for four different mental health facilities. Hall [1988] reports that both the distance to the centre and socio-demographic variables had a significant impact on the rate of utilization of the facilities, although the effect and significance of the variables on the utilization rate differed at each facility. Kanaroglou and Hall [1989] develop a nested logit model which examines both the probability of facility use and frequency of facility use and apply this model to the same data set. The results from the logit model, in general, confirm the same trends as the regression analysis conducted by Hall [1988].

While broad consensus has arisen over the factors that affect health care utilization in the developed world, no such consensus exists in developing countries, where health care systems are often "patchy, pluralistic, and under pressure" [Bailey and Phillips, 1990, p. 1]. However, Hellen [1986] points out that, in this context, many of the methods for health systems research used in developed countries, such as those relating to accessibility and optimization, could be adapted for use in developing countries.

One interesting study, conducted by Paul [1992] in a rural area of Bangladesh, examines the health-seeking behaviour of parents whose child had a fatal illness. A sample of 1800 women who had been or were married was selected from the study area. These women had a total of 152 children in the 1-4 age group die over a seven year period. Paul fitted a step-wise logistic regression model to the data examining the utilization of

qualified western doctors. From this model he found that the significant factors influencing utilization were sex of the child, distance from a doctor, and whether the child was first-born. Socio-economic factors were not a deterrent to health-seeking but the probability of seeking medical intervention from a qualified doctor exhibited a strong gender bias – it was 34% for girls and 66% for boys. Paul recommended both the expansion of facilities to increase geographic accessibility and a public health education program to discourage the gender biases of the health-seeking behaviour of the parents.

In fact, studies in developing countries often have conflicting results. For example, in a study in the Bicol region of the Phillipines, Akin *et al.* [1985, p. 162] find that "education, urban residence, and the perceived seriousness of illnesses" were the most important factors in determining health care utilization patterns. Chernichovsky and Meesok [1986] emphasize the effects of economic considerations on health care utilization in Indonesia. Another study in Grenada [Poland *et al.*, 1990] finds that in addition to age, gender, and mobility, variables that "act as proxies for underlying relationships based on health attitudes and behaviour (p. 23)" are important determinants of utilization.

Habib and Vaughn [1986] examine the effects that age, gender, nature of sickness, income and distance to the nearest health centre have on the utilization rates of health services in Iraq using a linear regression model. They found that the primary determinants in utilization were perceived sickness and the distance to nearest health centre.

As well as studies of accessibility to health care and primary health care, there are also some studies that specifically examine the accessibility of maternal-child health and family planning services. Since this service forms the basis of the empirical analysis presented in Chapter 7, these studies are discussed in the next section.

2.1.6 Accessibility of MCH/FP services

Although there have been numerous studies examining the potential accessibility of primary health care in developing countries, there are fewer studies that examine exclusively the potential accessibility of a target population to maternal-child health and family planning services. Partly, this is because the World Health Organization strongly encourages the integration of these services within primary health care provision [Hart *et al.*, 1990]. For example, Ayeni *et al.* [1987] examine the distance to the nearest maternal-child health centres as well as other primary health care facilities.

Rosero-Bixby [1993; 1995] uses the measure proposed by Joseph and Bantock [1982] to examine the accessibility of family planning services in Costa Rica. Also, many authors examine the utilization rates of maternal-child health along with utilization of primary health care. For example, the studies by Akin *et al.* [1985], Chernichovsky and Meesok [1986], and Habib and Vaughn [1986], mentioned in the previous section, also examine maternal-child health and come to the identical conclusions for this service as they did for primary health care.

Many factors have been found to influence utilization of contraception and family planning in developing countries. For example, the demand for contraceptives is dependent on cultural, religious, demographic, as well as psychological factors, such as the desire to limit or space births [Easterlin *et al.*, 1988]. The selective availability of methods from a large number of different providers, both medical and non-medical, can lead to complicated interactions between the types of provider and the types of methods used [Tsui and Ochoa, 1992]. This makes it difficult to discern the importance of the different factors affecting utilization. However, it is clear from general demographic trends that there is a large unmet need for contraception. The World Bank [1993] estimates that between 10 and 40% of married women of reproductive age in most developing countries have an unmet need, and that filling this need would reduce the fertility rate in most developing countries outside of Sub-Saharan Africa to near two children per woman.

Several studies have been conducted on family planning accessibility. The effects of distance to service providers on contraceptive utilization behaviour varies among studies and are often quite weak [Chen *et al.*, 1983; Cornelius and Novak, 1983; Tsui *et al.*, 1981; Tsui, 1982]. For example, in a study of distance deterrence on contraception usage in rural Bangladesh, Paul [1991] found no significant effect of distance on rates of both clinical and non-clinical contraceptive usage. However, when the interaction effects between distance to a paved road and distance to a facility are incorporated in the models, this interaction effect is statistically significant. Similarly, a recent study in Thailand [Entwisle *et al.*, 1995] using a geographic information system finds that travel time (estimated from a road network) is a good predictor of contraceptive utilization. The study also finds a "lagging" effect between the opening of a new facility and utilization behaviour changes.

2.1.7 Summary of Health Care Accessibility

Accessibility, both potential and realized, to health care is the product of a complicated inter-relationship between many different types of factors. The importance of these factors varies greatly from study to study. This reflects the diversity in the organization of various health systems and in the characteristics of the target populations [Joseph and Phillips, 1984]. Moreover, Phillips [1990] notes that "factors emphasized in studies of utilization tend to vary from one academic discipline to another (economists stressing cost factors; psychologists, various behavioural and socio-psychological matters, for instance)" (p. 195).

Despite these differences, a great deal of commonality exists between these different studies. Specifically, accessibility is viewed as a property of the interaction between the target population and the services that are available to them. Intervening factors, which either increase or decrease accessibility, can be classified into three main categories. The first category are characteristics of the population at risk. The second relates to the characteristics of the service delivery system and the last relates to the distance or spatial separation between the target population and the service delivery system. However, in all cases, in its most basic conceptualization, accessibility measures assess the opportunities for interaction between the service providers and a target population. This can either be the potential for interaction in the case of potential accessibility or the actual level of interaction for realized accessibility.

The next section discusses spatial interaction models. These models quantify the aggregate level of interaction between a set of origins and destinations based on the principles of entropy maximization. The models relate the level of interaction to the same three groups of characteristics: the origins (or target population), destinations (or service delivery system), and the distances (or spatial separations) noted above. Thus, they have a close link to models of health care accessibility, particularly of potential geographic accessibility.

2.2 Spatial Interaction Models

Spatial interactions models quantify the flows or levels of interaction from an area or group of areas to various destinations in a system. Broadly defined, a spatial interaction can be considered as movement or communication over space that is the result of

a decision process [Fotheringham and O'Kelly, 1989]. Spatial interactions are modelled within the conceptual framework of a system which has a number of interacting origins and destinations. Origins have a set of propulsiveness characteristics associated with them; destinations have a set of relevant attributes that influence their attractiveness. Finally, a distance or spatial separation is defined between each origin and destination. Thus the flows or levels of interactions between each origin-destination pair are a function of the constraint of distance, the attraction of increased opportunities, and the demand or propulsion of the origins.

Within this general framework, there are several different families of spatial interaction models which emphasize the importance of different constraints and the type of information desired. This is a result of their application in diverse areas such as retail shopping [Openshaw, 1973], migration [Ewing, 1976], and health care [Mayhew and Leonardi, 1982] among others. In the following sections, a mathematical framework for interaction models is provided first. Next, four different families of models are described. Then, several extensions to the basic spatial interaction model and the relationship between the models and accessibility measures are discussed. Finally, several examples of the models applied to health care planning are described.

2.2.1 A Mathematical Framework

Consider a system having N_O origin nodes interacting with N_D destination nodes. Each origin has N_P propulsiveness variables associated with it and each destination has N_A attractiveness variables. Let X_{ik}^P be the k^{th} propulsiveness variable for origin node i, X_{jk}^A be the k^{th} attractiveness variable for destination node j, D_{ij}^k be the k^{th} variable representing distance[1] between origin i and destination j, and T_{ij} be the level of interaction or flow[2] between the origin i and destination j. Define $T = [T_{ij}]_{N_O \times N_D}$ to be the matrix of flows. As well, it is convenient to define the following variables relating to the total

[1]It should be emphasized that distance, in this context, is used in a generic sense that describes the difficulty or impedance of travelling from origin i to destination j. Some possible measures of "distance" include travel time or cost of travel as well as other more traditional measures such as straight line (Euclidean) distance.

[2]Flow is used as a synonym for level of interaction.

flow for each node and for the system:

$$\begin{aligned} O_i &= \sum_j T_{ij} && \text{the total outflow from origin node } i, \\ I_j &= \sum_i T_{ij} && \text{the total inflow into destination node } j, \text{ and} \\ T_T &= \sum_{i,j} T_{ij} && \text{the total flow in the system.} \end{aligned} \tag{2.9}$$

Note that $T_T = \sum_i O_i = \sum_j I_j$. The goal of a spatial interaction model is to relate the flows between the various origins and destinations nodes, *i.e.*, to relate T to X^P, X^A, and the spatial separation.

The propulsiveness variables influence, either positively or negatively, the outflow from each origin. The selection of the relevant propulsiveness variables is dependent on the system under consideration. For example, the number of people in an area is a propulsiveness variable with a positive influence on interaction. Similarly, the attractiveness variables influence, either positively or negatively, the inflow into each destination and could include the size of, or resource availability at, a facility or the average waiting time. In this case, the attractiveness of a facility would increase with increasing facility size while the increased waiting time would have a negative impact.

For simplicity of notation only one attractiveness variable, one propulsiveness variable, and one distance variable are used in the following discussion. However, the results can be easily generalized to multivariate cases.

The D_{ij} values reflect the distance between the origin and destination nodes. The distance can be measured either subjectively or objectively. The subjective spatial separation between an origin and a destination reflects the separation or cost viewed by a subject at the origin. The subjective approach determines the distance by surveying individuals or households in order to determine their travel behaviour and preferences. Cadwallader [1975] uses the latter approach in a simple interaction model examining the patronage of supermarkets. Unfortunately, the subjective approach is not commonly used since it is both time-consuming and data-intensive. However, this approach is important where people's choice of a service facility is influenced by their knowledge of the available facilities. The objective approach uses measures such as distance, travel time, or travel cost, with distance being the most widely used measure [Fotheringham and O'Kelly, 1989].

Mathematically, a spatial interaction process can be expressed as

$$T_{ij} = f(\alpha_P, X_i^P, \alpha_A, X_j^A, \beta, D_{ij}) \tag{2.10}$$

where α_P, α_A, and β are parameters relating the value of the variables to the flow. There is "virtual unanimity of opinion" [Fotheringham and O'Kelly, 1989, p. 10] that the relationship between the attractiveness and propulsiveness variables and the flow is best modelled as a power function, *i.e.*

$$f_i^P = f_P\left(\alpha_P, X_i^P\right) = \left(X_i^P\right)^{\alpha_P} \qquad \text{and} \qquad f_j^A = f_A\left(\alpha_A, X_j^A\right) = \left(X_j^A\right)^{\alpha_A} \tag{2.11}$$

where f_i^P is the propulsiveness of origin i and f_j^A is the attractiveness of destination j.

On the other hand, there is not the same unanimity in agreement on the functional form for the deterrent effect of distance. The two forms that dominate the literature [Fotheringham and O'Kelly, 1989], as noted in Section 2.1.2, are

the power function	$f_D(\beta, D_{ij}) = (D_{ij})^{-\beta}$	(2.12)
and the exponential function	$f_D(\beta, D_{ij}) = \exp(-\beta D_{ij})$.	(2.13)

One difficulty of using the power function is that it tends to infinity as distance approaches zero, while the exponential function tends to one. This behaviour can be problematic in some systems, for example where the origin nodes and destination nodes are the same. As well, there are several alternative formulations for the distance decay function. For instance, Luoma and Palomäki [1983] propose a general distance decay function that has power and exponential functions as special cases.

2.2.2 Families of Spatial Interaction Models

Wilson [1974] defines four basic families of spatial interaction models that are differentiated by whether data on the inflow to the destination nodes and outflow from the origin nodes are exogenously defined. In the unconstrained model both attractiveness and propulsiveness variables are used. In production-constrained models the propulsiveness variables, X_i^P, are ignored and replaced with a vector of outflows for each origin node, O_i, while in attraction-constrained models the attractiveness variables, X_j^A, are

replaced with the inflows for each destination, I_j. Finally, in the production-attraction-constrained model[3] both the attractiveness and propulsiveness variables are replaced with their inflows and outflows respectively.

Choukron [1975] notes that there are numerous possible derivations for spatial interaction models. One formulation suggests deriving the model by aggregating decision processes at the individual level. This approach can be embedded within the random utility framework discussed in the next chapter in the context of a generic model for accessibility. An alternative approach, suggested by Wilson [1974] among others, is to develop spatial interaction models using the concept of entropy maximization. However, information minimization can also be used so that the flows between nodes in the system are calculated by finding the most probable distribution of flows given the observed values of the attractiveness, propulsiveness, and spatial separation variables. This distribution is given by a configuration that minimizes the Kullback information gain (KIG) associated with choosing the distribution of flows [Snickars and Weibull, 1977]. The KIG is given by

$$KIG = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.14}$$

where p_{ij} is the probability of an individual in zone i interacting with zone j and q_{ij} is the prior information on this interaction. If p^*_{ij} minimizes (2.14), then this is the assignment which has the lowest information content. With any other assignment of probabilities some private information bias has been added which is not justifiable in terms of known information about the system [Snickars and Weibull, 1977]. A complete treatment of the use of information measures in a spatial context is provided by Webber [1979]. In this context p_{ij} is given by T_{ij}/T_T. The q_{ij} values are given by previous information. This could be interaction values from a previous time period [Snickars and Weibull, 1977], probabilities based on the likely occurrence of different events [Webber and O'Kelly, 1981], or the values of site-specific variables that incorporate prior knowledge of the origins and destinations [Fotheringham and O'Kelly, 1989]. It should be noted that both the p_{ij} and q_{ij} values in (2.14) are properly expressed as proportions.

The four families of spatial interaction models described by Wilson [1974] may be derived in terms of an optimization problem. The propulsiveness and attractiveness

[3]Also referred to as the doubly-constrained model.

variables are included as prior knowledge about the system and the multiplicative constants are removed from the objective since it is invariant to scaling. The optimization problem can be formulated as follows.

$$\underset{T_{ij}}{\text{Minimize}} \sum_{i,j} T_{ij} \log\left(\frac{T_i j}{f_i^P f_j^A}\right) \tag{2.15}$$

subject to

$$\sum_j T_{ij} = O_i \qquad i = 1 \ldots N_O \tag{2.16}$$

$$\sum_i T_{ij} = I_j \qquad j = 1 \ldots N_D \tag{2.17}$$

$$\sum_{i,j} T_{ij} D_{ij} = D_T. \tag{2.18}$$

Depending upon the family of model, either zero, one, or both of the constraints (2.16) and (2.17) may hold. The distance constraint (2.18) causes the model to reproduce some total trip length D_T. The different families of spatial interaction models are now generated from the solution to this optimization problem using the framework proposed by Wilson [1974] that was generalized using the Kullback information gain by Fotheringham and O'Kelly [1989].

Unconstrained Models Since in the unconstrained model neither the inflow nor outflow totals are known, the problem becomes the minimization of (2.15) subject only to the total trip length constraint (2.18). The result of this minimization is

$$T_{ij} = f_i^P f_j^A \exp(-\beta D_{ij}) \tag{2.19}$$

where β is chosen to satisfy the total trip length constraint. Fotheringham and O'Kelly [1989] note that these unconstrained models are not particularly useful since the information they provided is generally of poor quality.

Production-Constrained Models In this type of model, the flows leaving each origin are known. The information on the propulsiveness of the origin is not used and the propulsiveness values, f_i^P, are set to an arbitrary and equal value, say 1. The objective (2.15) is minimized subject to (2.16) and (2.18). The result of this optimization is

$$\begin{aligned} T_{ij} &= B_i^{-1} f_j^A O_i \exp(-\beta D_{ij}) \\ B_i &= \sum_j f_j^A \exp(-\beta D_{ij}). \end{aligned} \tag{2.20}$$

The values of B_i ensure that the total flow out of an origin node is reproduced by the model and is referred to as the balancing factor.

Attraction-Constrained Models In this case, the flows entering each origin are known. The attractiveness variable, f_i^A, is set to an arbitrary constant value and the objective (2.15) is minimized subject to the destination node inflow constraints (2.17) and the total distance constraint (2.18). This results in

$$\begin{aligned} T_{ij} &= C_j^{-1} f_i^P I_j \exp(-\beta D_{ij}) \\ C_j &= \textstyle\sum_i f_i^P \exp(-\beta D_{ij}). \end{aligned} \tag{2.21}$$

Again, the C_j variables balance the flow so that the total flow entering a destination node equals the exogenously defined value I_j.

Attraction-Production-Constrained Models In doubly-constrained models both the outflow from each origin node and the inflow to each destination node are known. Therefore the objective (2.15) is minimized subject to (2.16), (2.17), and (2.18). The resulting expression is

$$\begin{aligned} T_{ij} &= B_i^{-1} C_j^{-1} O_i I_j \exp(-\beta D_{ij}) \\ B_i &= \textstyle\sum_j C_j^{-1} I_j \exp(-\beta D_{ij}) \\ C_j &= \textstyle\sum_i B_i^{-1} O_i \exp(-\beta D_{ij}). \end{aligned} \tag{2.22}$$

Here B_i and C_j are interrelated balancing factors which ensure that the flow out of the origin nodes and flow into the destination nodes balance.

As well, there are some straightforward generalizations to the doubly-constrained model. Snickars and Weibull [Snickars and Weibull, 1977] suggest incorporating information on the interaction pattern from the previous time period. This leads to a solution of

$$\begin{aligned} T_{ij} &= B_i^{-1} C_j^{-1} O_i I_j S_{ij} \exp(-\beta D_{ij}) \\ B_i &= 1/\textstyle\sum_j C_j^{-1} I_j S_{ij} \exp(-\beta D_{ij}) \\ C_j &= 1/\textstyle\sum_i B_i^{-1} O_i S_{ij} \exp(-\beta D_{ij}) \end{aligned} \tag{2.23}$$

where S_{ij} is the flow between origin i and destination j in the previous time pe-

riod.

It should be noted that there has been concern over model misspecification and the interpretation of the distance-decay measures [Fotheringham, 1981; Fotheringham, 1983; Fik and Mulligan, 1990]. In particular, Fotheringham [1981] notes that there is often a spatial pattern evident when a gravity model is calibrated with origin-specific distance-decay parameters. Several extensions of the standard gravity model have been proposed to address these difficulties. These extensions include the Alonso framework [1978], relaxed models [Halleford and Jörnsten, 1985], changing masses model [Luoma and Palomäki, 1983], and the Tobler model [1983].

2.2.3 Linkages to Measures of Accessibility

Leonardi [1978] notes that there is a strong linkage between spatial interaction models and several potential geographic measures of health care accessibility. As noted previously, health care accessibility is a property of the interaction between the target population for health services and service providers. The goal of spatial interaction models is to quantify the level of these interactions and the effects that explanatory variables have upon them.

Suppose that the propulsiveness of an origin or population group is defined as its total population, *i.e.*, $f_i^P = P_i$ and that the attractiveness of the destination or facility is defined by its size so that $f_j^A = S_j$. The standard Hansen [1959] measure of accessibility can be expressed as

$$A_i = \sum_j S_j f_{ij}^D = (1/P_i) \sum_j P_i S_j f_{ij}^D = (1/P_i) \sum_j T_{ij} \tag{2.24}$$

where T_{ij} is defined as in the unconstrained model (2.19). The accessibility measure proposed by Joseph and Bantock [1982] also has the same interpretation. In this instance,

$$\begin{aligned} A_i &= \sum_j (S_j/C_j) f_{ij}^D = (1/P_i) \sum_j P_i (S_j/C_j) f_{ij}^D = (1/P_i) \sum_j T_{ij} \\ C_j &= \sum_i P_i f_{ij}^D . \end{aligned} \tag{2.25}$$

This is identical in form to the attraction-constrained spatial interaction model (2.21)

with $I_j = S_j$. In both of these models, accessibility is defined as

$$A_i = \sum_j T_{ij}/P_i \tag{2.26}$$

so that the accessibility is equivalent to the per capita level of interaction. Minimizing the overall variance of (2.25) is an objective of the model proposed by Mayhew and Leonardi [1982]. This is discussed in the next subsection.

2.2.4 Application to Health Care Systems

Several authors have applied spatial interaction models to examine both the accessibility and the allocation of resources in health care systems, but these studies have been applied in developed countries only. Despite an extensive literature search, no published research could be found where spatial interaction models have been applied to modelling health care accessibility in developing countries.

Martin and Williams [1992] applied a spatial interaction model at the level of primary health care in England using population disaggregated to a 100 metre grid. First, an attraction-constrained spatial interaction model was calibrated to patient registration data. This model was then used to assess the accessibility within the study area using several different measures such as the minimum distance, the mean distance, the Hansen accessibility measure, and the log-sum measure ($(1/\beta)\ln\sum_j f_j^A \exp(-\beta D_{ij})$). This allowed for a sophisticated examination of the differential accessibility within a region. In addition, they demonstrated that the framework allows for the estimation of market-catchments for each practice. Furthermore, a subsequent paper [Martin *et al.*, 1994] examined these issues at a regional scale to determine the allocation of deprivation payments to GP practices.

Mayhew and Leonardi [1982] introduced a framework for examining the allocation of resources at the urban and regional level. Their model was based on an attraction-constrained spatial interaction model (2.21) with f_i^P being the propensity of area i to generate patients. The decision variables[4] s_j were the case-load capacity of destination region j. Mayhew and Leonardi then formulate nonlinear models for optimally allocating resources (s_j) among the destination zones according to four criteria: equity, efficiency, and two distance measures. The amount of resources allocated to each zone was

[4]Decision variables are expressed in lower case letters.

only allowed to change by a certain proportion so that the s_j values were constrained to lie between minimum and maximum values. These models are briefly presented below.

Equity The first model is based on the equity criterion and attempts to allocate s_j so that the number of patients generated in each zone is proportional to its relative need. This can be formulated as the following optimization problem.

$$\underset{s_j}{\text{Minimize}} \sum_i \left(\sum_j (s_j/C_j) f_{ij}^D - \overline{A} \right)^2 \tag{2.27}$$

Subject to

$$S_j^{\min} \leq s_j \leq S_j^{\max} \quad j = 1 \ldots n \tag{2.28}$$

$$\sum_j s_j = Q \tag{2.29}$$

where Q is the total resources available in the system and $\overline{A} = Q / \sum_i f_i^P$. Note that the second term in the objective function reflects the total resources in the system per unit of relative need while the first term is simply the Joseph and Bantock [1982] accessibility measure, the total usage of the system from this zone per unit of relative need. Thus, the goal of this model is to reduce the variation in accessibility or resource availability of each zone.

Efficiency The efficiency criterion allocates resources in the system so that patient preferences for places of treatment are maximized. This objective can be thought of as preferentially allocating resources to areas that have a large potential demand on them, *i.e.*, zones where C_j is large. Mathematically this is expressed as

$$\underset{s_j}{\text{Maximize}} \; -\sum_j s_j \left[\log (s_j/C_j) - 1 \right] \tag{2.30}$$

subject to (2.28) and (2.29) as defined previously.

Distance 1 This criterion chooses a resource configuration that attempts to equalize the average distance from places of residence to places of treatment. If the average distance of origin zone i is defined as

$$D_i = \frac{\sum_j T_{ij} D_{ij}}{\sum_j T_{ij}} = \frac{\sum_j I_j C_j^{-1} D_{ij} f_{ij}^D}{\sum_j I_j C_j^{-1} f_{ij}^D} \tag{2.31}$$

and the system average accessibility cost is defined as

$$\overline{D} = \frac{\sum_i \sum_j T_{ij} D_{ij}}{T_T} \tag{2.32}$$

then the objective for this minimization problem may be formulated as

$$\underset{s_j}{\text{Minimize}} \sum_i \left(D_i - \overline{D}\right)^2 \tag{2.33}$$

subject to (2.28) and (2.29) as defined before.

Distance 2 Another distance criterion is to minimize the variance in the distances since it may be difficult to equalize the average distance. The variance in distance is defined as

$$\frac{\sum_j T_{ij} \left(D_{ij} - \overline{D}\right)^2}{\sum_j T_{ij}} = \frac{\sum_j (D_{ij} - \overline{D})^2 C_j^{-1} f_{ij}^D s_j}{\sum_j C_j^{-1} f_{ij}^D s_j}. \tag{2.34}$$

The objective can then be formulated as

$$\underset{s_j}{\text{Minimize}} \sum_i \frac{\sum_j (D_{ij} - \overline{D})^2 C_j^{-1} f_{ij}^D s_j}{\sum_j C_j^{-1} f_{ij}^D s_j} \tag{2.35}$$

subject to (2.28) and (2.29).

Mayhew and Leonardi [1982] applied their models on 1977 data for the London region in England using thirty-three origin zones (administrative boroughs of the Greater London District) and thirty-six health districts as the destination zones with one external zone to close the system. Their testing indicated that both of the distance criteria produced unacceptable results. On the test data, the Distance 1 measure allocated resources to the least accessible zones so that the population had the same, albeit poor, accessibility, and the Distance 2 measure showed unpredictable behaviour in sensitivity tests. Based upon this, they develop a multiobjective model for resource allocation combining equity and efficiency objectives to allow the planner or relevant decision maker to examine the trade-offs between efficiency and equity.

Taket [1989] uses a spatial interaction model to examine the accessibility and equity in the future provision of in-patient hospital facilities in East Anglia, England. In his

study, Taket formulates an attraction-constrained spatial interaction model using the same variables as in Mayhew and Leonardi [1982], namely f_i^P being the propensity of area i to generate patients, and S_j being the case-load capacity in zone j. This model was calibrated using data for 1981 and validated with data for 1985 at the Local Authority level and used to project the relative need in each area for the year 2001.

Three scenarios for the allocation of resources for the year 2001 were evaluated. The study calculated a measure of equity for each region, defined as $\sum_j T_{ij}/P_i$ (corresponding to the Joseph and Bantock accessibility measure) and the average distance (which was termed accessibility) for each region, $D_i/\overline{D}$ where D_i and $\overline{D}$ are as defined previously. Taket [1989] found that the most decentralized scenario provided the best results in terms of both equity and average distance. It is interesting to note that, as opposed to the previous study which optimized the distribution of resources in terms of different criteria, this study examines three different resource allocation scenarios and evaluates them using two different criteria.

Wilson and Gibberd [1990] follow a similar approach in developing a multiobjective model to allocate resources in a regional health care system according to four criteria: minimizing operating cost, minimizing transportation costs, equalizing utilization with respect to relative need, and equalizing accessibility. An interesting aspect of their model is the inclusion of a method to examine the dynamic aspects of reallocating resources since it is impractical to make large changes in resource allocations immediately and the relative need in each zone is a function of time. After formulating the dynamic problem in the form of a differential equation with future costs discounted using exponential functions, this equation was solved in order to obtain the optimal change in resource allocation for each time period for an appropriate parameter value. They give an example application using data for New South Wales, Australia.

In a series of papers, Segall [1988; 1989a; 1989b] applies and extends the model proposed by Mayhew and Leonardi [1982] using hospital utilization information for Massachusetts. The extensions include: partitioning the destination zones into a two-level hierarchy, examining the effect of closing hospitals, and disaggregating the patient flows by treatment type. As well, Segall derives a production-constrained version of the models and objectives discussed in Mayhew and Leonardi, as well as a stochastic version of the production-constrained model, although these were not tested using real data.

2.2.5 Summary of Spatial Interaction Models

Spatial interaction models are based on a framework that interrelates origin node outflows and propulsiveness variables, destination node inflows and attractiveness variables, and the spatial separation between origin and destination nodes.

Four families of spatial interaction models were summarized in this section. The unconstrained model corresponds to the case where neither outflows from the destination nodes nor inflows to the destination nodes are constrained. In the production-constrained model, the outflow is constrained. The inflow to each destination node is constrained in the attraction-constrained model. Finally, in the production-attraction-constrained model both the inflow and the outflow are constrained.

Spatial interaction models have been applied to flows of hospital patients in order to examine the accessibility, equity, and efficiency of resource allocations within a health care system. As well, recent work has applied spatial interaction models to evaluate the accessibility of primary health care.

The next section discusses facility location models. In contrast to a spatial interaction model, which assumes that the destinations are fixed, facility location models attempt to determine the optimal configuration of facilities for a given demand configuration.

2.3 Facility Location Models

In general, facility location problems can be divided into three main groups, based on the type of restrictions that are placed on the location of new facilities [Hansen *et al.*, 1987]. In continuous location problems, the set of possible new locations is limited to a subset of the plane. If the new facility locations are constrained to lie along a network, the problem is known as a network location problem. Finally, if the problem is to choose optimal facility locations from a finite set of candidate facility sites chosen by some prior analysis, the problem is known as a discrete location problem.

The focus of this discussion is on discrete facility location problems since the vast majority of facility location models that have been applied to health care planning in developing countries are discrete (for example, [Eaton *et al.*, 1981; Mehretua *et al.*, 1983; Tien and El-Tell, 1984; Ayeni *et al.*, 1987; Oppong, 1992]). In addition, compared to continuous problems, discrete models allow for considerable flexibility in the specification

of the characteristics of the facilities, the target population, and the distances without dramatically altering the model structure and solution techniques [Hansen *et al.*, 1987].

This section provides an overview of the basic aspects of discrete facility location problems. The first part of the section discusses several different classes of discrete facility location problems presenting both mathematical formulations and solution methods. These models assign all the users associated with a demand node to the nearest open facility. The use of these models for health care planning in developing countries is also reviewed. As noted in Section 2.1.2, nearest-centre allocation may not hold for real-world travel patterns. Therefore, alternative models are discussed where the nearest-centre allocation rule is relaxed. Users choose the facilities they attend subject to a distance decay effect that combines aspects of classical facility location theory with the spatial interaction models previously discussed.

2.3.1 Discrete Facility Location Models with Nearest-Centre Allocation

Facility location models are used for both determining locations of new service providers and for comparing the efficiency of the current spatial configuration of the system to an optimal configuration. The majority of research on these models have used one of two main families of well-known discrete facility location models, based on whether efficiency or equity is the primary objective.

The objective of a site-selecting facility location problem is to locate new facilities to serve an existing set of users. These new facilities are selected from a finite set of candidate locations determined through previous analysis. If there are existing facilities, these can be incorporated in the problem by ensuring that the existing facilities are included in the candidate subset. The various forms of the discrete facility location problem with the users allocated to the nearest facility are now considered.

The generic discrete facility location problem can be defined as follows. Let N_O be the number of users and N_D be the number of candidate facility sites. Any locational alternative can be represented by a binary vector $\mathbf{y} = [y_1, y_2, \ldots, y_{N_D}]$ where the decision variable y_j is defined as

$$y_j = \begin{cases} 1 & \text{if a new facility is located at site } j \\ 0 & \text{otherwise.} \end{cases}$$

In addition, certain classes of problems require consideration of the allocation of demand to facilities. Let x_{ij} be the proportion of the demand from user i allocated to candidate facility site j, and $\mathbf{x} = [x_{11}, \ldots, x_{1,N_D}, x_{21}, \ldots, x_{N_O,N_D}]$ represent a vector of these allocations. Finally, define $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ as the decision vector defining a solution to the problem.

If $\mathcal{A}$ is the set of feasible locational alternatives and allocations to the location problem, then a single criterion location problem can be formulated as follows where $Z(\cdot)$ is the objective function.

$$\text{Minimize } Z(\mathbf{z}) \tag{2.36}$$

$$\text{subject to } \mathbf{z} \in \mathcal{A}. \tag{2.37}$$

The subsequent formulations make use of the following additional definitions. Let P_i be the demand associated with user i and D_{ij} be the transportation cost for one unit of demand from user i to site j.

The following discrete facility location problems are considered: the p-median problem, the uncapacitated facility location problem, covering problems, the p-centre problem, and the hierarchical facility location problem. Mathematical formulations and solution methods are discussed for each type of problem. Finally, there is a discussion of the application of these models to health care planning in developing countries.

The p-Median and Uncapacitated Facility Location Problems

The objective of the p-median problem is to locate p facilities so that the total transportation cost is minimized [Hansen *et al.*, 1983]. The p-median problem can be formulated as an integer linear programming problem as follows.

$$\underset{\mathbf{x},\mathbf{y}}{\text{Minimize}} \sum_{i,j} P_i D_{ij} x_{ij} \tag{2.38}$$

subject to

$$\sum_j x_{ij} = 1 \qquad i = 1, \ldots, N_O \tag{2.39}$$

$$0 \leq x_{ij} \leq y_j \qquad i = 1, \ldots, N_O \quad j = 1, \ldots, N_D \tag{2.40}$$

$$y_j \in \{0, 1\} \qquad j = 1, \ldots, N_D \tag{2.41}$$

$$\sum_j y_j = p. \tag{2.42}$$

In this formulation, constraint (2.42) ensures that exactly p new facilities are located.

Constraint set (2.39) ensures that the total demand from each user is satisfied while constraint set (2.40) ensures that no demand is satisfied at site j if a facility is not established there. Constraint set (2.41) is the integrality restriction related to the definition of variable y_j. In an optimal solution, the x_{ij} values will all be either one or zero [Hansen *et al.*, 1983]. Finally, it is possible to develop a simplified formulation of the p-median problem by dropping the allocation variables. Define the function

$$D(i, \mathbf{y}) = \min_{j, y_j = 1} D_{ij} \tag{2.43}$$

to measure the distance to the nearest selected facility site. With this definition, the p-median problem can be formulated as follows.

$$\underset{\mathbf{y}}{\text{Minimize}} \sum_i P_i D(i, \mathbf{y}) \tag{2.44}$$

subject to

$$y_j \in \{0, 1\} \quad j = 1, \ldots, N_D \tag{2.45}$$

$$\sum_j y_j = p. \tag{2.46}$$

Note that this revised formulation is no longer an integer linear program.

The p-median problem is an NP-hard problem [Kariv and Hakimi, 1979b], which implies that no exact polynomial-time algorithm is known for this type of problem [Johnson and Papdimitrou, 1985]. One solution approach for the p-median problem involves relaxing the constraint that y_j be zero or one (2.41) and replacing it with

$$0 \leq y_j \leq 1. \tag{2.47}$$

This gives a linear programming problem whose solutions are often integral [Hansen *et al.*, 1987] and whose objective is always a lower-bound to the optimal value of the previous problem. However, the direct solution of this problem is difficult due to the $N_D(N_O + 1)$ variables and $(N_D + 1)(N_O + 1)$ constraints [Hansen *et al.*, 1983]. This can be overcome through the use of Lagrangian relaxation as proposed by Narula *et al.* [1977]. The constraints defined by (2.39) are associated with Lagrangian multipliers, λ_i, and the objective function (2.38) is replaced with

$$\underset{\mathbf{x}, \mathbf{y}, \lambda_i}{\text{Minimize}} \sum_{ij} \left(P_i D_{ij} - \lambda_i \right) x_{ij} + \sum_i \lambda_i \tag{2.48}$$

subject to constraints (2.40), (2.42), and (2.47). If the λ_i values are fixed, the optimal values of x_{ij} can be obtained as $x_{ij}^* = y_j$ when $P_i x_{ij}^* - \lambda_i \leq 0$ and $x_{ij}^* = 0$ otherwise. The values of y_j can be found by selecting those sites corresponding to the p smallest values of $\sum_i b_i x_{ij} - \lambda_i$ [Hansen *et al.*, 1983]. The values of the multipliers, λ_i, can be determined through a subgradient search method to maximize the dual of this problem [Narula *et al.*, 1977].

Several different heuristic strategies have been proposed for the p-median problem that involve the addition or deletion of one new facility or the interchange between a currently selected facility and a vacant facility site. These heuristics are flexible and can be applied to a variety of extensions and variations of the standard problem [Leonardi, 1983].

In the Drop heuristic, first proposed by Feldman *et al.* [1966], facilities are initially located at every demand site j. At each iteration, the facility that causes the least increase in total cost when dropped is eliminated. The process is repeated until there are only p facilities remaining. The Add method was first proposed by Kuehn and Hamburger [1963]. With this method, one facility is located at the site of least total cost. In each iteration, the facility whose addition causes the greatest decrease in total cost is added. The iterations are continued until exactly p facilities have been located.

In contrast to the two previous heuristics, the Interchange procedure, initially proposed by Teitz and Bart [1968], operates on an existing pattern of p facilities. At each iteration, a single facility is moved to a vacant site as long as this causes a decrease in total cost. When there are no possible moves left that cause such a reduction, the procedure terminates. Densham and Rushton [1992] suggest some algebraic procedures for making the Interchange procedure operate effectively and efficiently on large scale problems in a microcomputer-based environment.

Numerous extensions have been proposed to the p-median problem. For example, Toregas *et al.* [1971] add a maximum travel cost constraint so that all users are within a specified travel cost or distance of a facility. This can be accomplished by setting $D_{ij} = \infty$ for the cases where this constraint is violated. Other extensions include adding capacity constraints, budgetary constraints, distance-sensitive demand, and hierarchical facility systems [Hansen *et al.*, 1983]. Hillsman [1984] generalizes the p-median problem into a unified linear model which can incorporate many different extensions.

A related problem is the uncapacitated facility location problem. As opposed to

p-median problem, the number of new facilities is not specified. Instead, a cost, E_j is associated with opening a facility at site j and the model objective is to minimize the total cost of the system, *i.e.*, the cost of establishing the facilities and the transportation cost. Mathematically this can be formulated as follows.

$$\underset{x_{ij},\, y_j}{\text{Minimize}} \sum_{i,j} P_i D_{ij} x_{ij} + \sum_j E_j y_j \tag{2.49}$$

subject to constraints (2.39), (2.40), and (2.41). Like the p-median problem, the uncapacitated facility location problem is an *NP*-hard problem and similar solution strategies have been applied to it. Hansen *et al.* [1987] report that both the Add, Drop, and Interchange heuristics and the Lagrangian relaxation technique perform well on this problem.

The Covering and p-Centre Problems

Covering problems involve locating a set of facilities so that user i is within an exogenously defined range, R_i, of a facility. Define $\mathcal{R}_i = \{j | D_{ij} \leq R_i\}$ as the set of potential facility sites within range of user i. In one formulation, the costs of establishing the facilities are minimized. This problem is known as the set-covering problem and is specified as follows.

$$\underset{y_j}{\text{Minimize}} \sum_j E_j y_j \tag{2.50}$$

subject to

$$\sum_{j \in \mathcal{R}_i} y_j \geq 1 \qquad i = 1, \ldots, N_O \tag{2.51}$$

$$y_j \in \{0, 1\} \qquad j = 1, \ldots, N_D \tag{2.52}$$

where the summation is over the set of potential facility sites within range of user i. Constraint set (2.51) ensures that each client is "covered" by a facility and constraint set (2.52) restricts y_j to integral values.

The set-covering problem can often be transformed into a significantly smaller problem through the application of reduction rules which remove redundant sites [Love *et al.*, 1988]. The reduced problem can be solved by relaxing the integer constraint (2.52) as before and solving the problem as a linear programming problem. If the corresponding solution is not feasible then a branch-and-bound approach may be applied [Balas and

Ho, 1980].

If the number of facilities found by the set-covering problem is infeasible due to cost considerations, one can increase the acceptable ranges, R_i, or cover the maximum possible demand with p new facilities. This later problem is known as the maximum covering problem. A mathematical formulation of this objective is as follows.

$$\underset{x_{ij}, y_j}{\text{Maximize}} \sum_i \sum_{j \in \mathcal{R}_i} P_i x_{ij} \tag{2.53}$$

subject to the p-median problem constraints, namely (2.39), (2.40), (2.41), and (2.42). This problem can be easily transformed to the p-median problem by modifying the D_{ij} variables in the objective (2.38) and the standard p-median techniques may then be used on the transformed problem [Hillsman, 1984].

As opposed to the set-covering and maximum covering problems which use an exogenously-specified distance standard, the p-centre problem locates facilities so that the maximum distance between a user and the nearest facility is minimized. One interpretation of the p-median problem is that it minimizes the average travel cost in the system. However, this criterion does not examine the effect of the facility locations on an individual user. A corresponding objective that examines the equity in the system, in a limited sense, would be to minimize the maximum travel cost for *any* user of the system. This problem is known as the p-centre problem and can be mathematically formulated as follows.

$$\underset{x_{ij}, y_j}{\text{Minimize}} \max_{i,j} P_i D_{ij} x_{ij} \tag{2.54}$$

subject to (2.39) through (2.41), the p-median constraints. Note that in many situations each origin or demand node is equally weighted in which case the objective (2.54) is replaced with $\max_{i,j} D_{ij} x_{ij}$. As with the other problems, this problem is also *NP*-hard [Kariv and Hakimi, 1979a].

One solution method for this problem suggested by Minieka [1970] involves solving a sequence of set-covering problems. In this approach, an upper bound for the value of (2.54), R_U, is obtained from the solution of the p-median problem while a lower bound, R_L, is taken as zero. Defining $\overline{R} = (R_U + R_L)/2$, the set covering problem is then solved with $R_i = \overline{R}$. If the number of facility locations is less than p then R_L is

set to $\overline{R}$ otherwise R_U is set to $\overline{R}$. This process is then repeated until $R_U - R_L$ is within some desired tolerance.

Furthermore, the p-median problem can be combined with the p-centre problem. For example, one might solve a p-centre problem to find the minimum travel cost standard, r, for the system using p facilities. A p-median problem with a maximum allowable travel cost of R might then be solved to find an efficient placement of facilities [Hansen *et al.*, 1983].

Hierarchical Facility Location Models

Many service delivery systems, such as rural primary health care systems in developing countries, are hierarchically organized [Hodgson, 1988]. A hierarchical system is organized into N_L levels of facilities. The level of a facility in the hierarchy is defined by the highest order of good or service it provides. There are two types of facility hierarchies: successively-inclusive, and successively-exclusive [Narula, 1984]. In a successively-inclusive hierarchy, a facility of level k offers all services of order $k, \ldots, N_L$ so that a high order facility offers low and medium order services as well as high order services[5]. On the other hand, a successively-exclusive hierarchy offers services that are unique to it. Although, hierarchical facility location models have not been widely applied [Beaumont, 1987], most hierarchical facility location models for public services are based on a successively-inclusive facility hierarchy.

Hodgson [1984] defines a set-based formulation for the hierarchical facility location as follows. Define $\mathbf{y}_k$ as the vector of binary decision variables y_{jk} where y_{jk} is one if a facility of level k is located at candidate facility site j and zero otherwise. Let F_k be the proportion of usage of level k services where p_k is the number of facilities of order k to be located. This problem can be formulated as follows.

$$\underset{\mathbf{y}_1, \ldots, \mathbf{y}_{N_L}}{\text{Minimize}} \quad \sum_k F_k \sum_i P_i D(i, \mathbf{y}_k) \tag{2.55}$$

subject to

$$y_{j,k-1} \leq y_{jk} \qquad j = 1, \ldots, N_D \quad k = 2, \ldots, N_L \tag{2.56}$$

$$\sum_j y_{jk} = \sum_{\ell=1}^{k} p_\ell \qquad k = 1, \ldots, N_L \tag{2.57}$$

$$y_{jk} \in \{0, 1\} \qquad j = 1, \ldots, N_D \quad k = 1, \ldots, N_L.. \tag{2.58}$$

[5]Note that the higher level facilities and higher order goods and services are denoted with lower indices, *i.e.*, the highest level of facility is denoted by $k = 1$.

Constraint (2.56) defines this model to be a successively-inclusive model, since all higher order facilities are included in the set of lower-order facilities, while constraint (2.57) specifies the number of facilities to locate of a given order (and all higher orders).

Fisher and Rushton [1979] outline three level-by-level methods for solving the hierarchical problem:

1. *top-down*, where the highest level facilities are located first and lower-level facilities are constrained to include higher level facilities;
2. *bottom-up*, where the lowest level facilities are identified first and candidate facility sites for higher level facilities are a subset of the identified facilities of the previous level; and,
3. *middle-out*, a combination of the other two strategies, the middle level facilities are first located, higher level facilities are selected from them and lower level are constrained to include them.

Hodgson [1984] criticizes these stepwise approaches, as the service levels are not independent of each other. Consequently, he proposes a simultaneous approach, locating all facility levels at once, using a modified version of the Interchange heuristic. Using several test problems, he finds that the simultaneous approach performs much better than the top-down method and somewhat better than the bottom-up method. However, it is often not possible to use this approach with real world data as it requires knowledge of the relative usage of facilities by their level [Oppong, 1992].

2.3.2 Applications to Health Care Planning

Eaton *et al.* [1981] applied a maximum covering problem for the location of rural health centres in Colombia. This problem was solved using a combination of the Add and Interchange heuristics. The number of health centres located was varied and the percentage coverage of the population was examined. The results were compared to the sites chosen by local planners. Their results indicated that although 78% of the population was being covered by the current 24 sites, the same level of coverage could be provided by only 15 sites located using an optimization approach. Fully 90% of the population could have been covered if all 24 facilities had been optimally located. This

study is extended by Bennett *et al.* [1982] to determine the locations of new health facilities which would most improve population coverage. Logan [1985] notes that in Sierra Leone, locating rural health centres in the administrative centres is considerably more costly than a more dispersed location pattern determined through facility location models.

Mehretu *et al.* [1983] use a p-median model with a maximum distance constraint in order to locate primary health posts in the Eastern Region of Burkina Faso. Mehretu [1985] uses the same region and examines methods of equitably allocating additional facilities. The proposed allocation procedure first assigns new primary health posts to villages in sub-regions that are the most deprived of resources. Then additional resources are allocated to other communities. Reid *et al.* [1986] use a set-covering model to find the minimal number and optimal locations for depots to supply primary health centres in two provinces in Ecuador. Two different formulations were solved: one based on distance and another based on travel time. They find that the set-covering approach reduced the number of depots needed relative to the current system. Further, Eaton *et al.* [1986] use facility location models to determine optimal ambulance deployment patterns in Santo Domingo.

Tien and El-Tell [1984] apply a two-level variant of the p-median problem. This model located primary health care facilities to minimize the total distance travelled by users and attached village clinics to larger health centres so that physicians based at the latter were able to visit the attached clinics. The model was applied to sample data from Jordan and solved using a Lagrangian relaxation technique. The solutions to the model indicated that significant gains could be made in both clinic accessibility and physician availability over the current system with only moderate locational changes.

The multilevel approach has been applied by several other authors in the context of health care planning in developing countries. For example, Dökmeci [1979] presented a multiobjective facility location model for a four level regional health system in Turkey. This model is based on a p-median problem with the minimization of both travel and facility costs. As well, the utilization of the system was calculated by $\sum_{ij} P_i D_{ij}^{-\beta} x_{ij}$. The optimal locations were first calculated for the lowest level. The facilities located from this solution were defined as the users in the next level. This process was repeated until all the levels of facilities had been located.

Moore and Revelle [1982] also apply a multilevel approach in a maximum covering problem. In their model, the service range for a higher level facility was greater than

for a lower level facility and the objective was to locate these facilities to minimize the population that is not covered by the facilities. This model was applied to data from Honduras for locating medical facilities and the coverage of the population was examined according to various budgetary constraints.

Oppong [1992] takes a similar approach in the Suhum District, Ghana. He uses a three-level hierarchical p-median problem and calculates the average weighted distance to each type of facility. Furthermore, he compares the systems in both dry season conditions and rainy season conditions (with only facilities located on all-season roads open). He finds that the optimal dry season system performs almost as poorly as the actual system during the rainy season. If facilities are restricted so that they can be located only at sites with all-season access the average distance during the dry season is 17% more than optimal dry season system. However, during the rainy season the average distance is 25% less. Thus, Oppong [1992] notes that it is important to incorporate local conditions in the analysis.

Hodgson [1988] points out some of the limitations in applying p-median models to primary health care planning. These shortcomings include the assumption that all residents at a given location attend the nearest facility, and that accessibility varies linearly with distance. Similarly, for p-centre problems, Hodgart [1978, p. 27] notes that the solution may "inflict excessive travel on the majority in order to reduce travel for a few isolated users." Coverage models exhibit sensitivity to the value of the exogenously defined maximum range [Mulligan, 1991]. Changing the maximum range can lead to a completely different locational configuration of facilities. These difficulties are confounded in situations where demand is spatially aggregated. This aggregation can cause large errors in the solution and incorrect estimates of coverage levels [Current and Schilling, 1989].

Oppong [1992] points out further deficiencies of the hierarchical models. For example, these models fail to recognize the differential attractiveness of different types of facilities owing to the mix of services provided. Consequently, he notes that users often bypass lower level facilities for a variety reasons, such as the perception of better service at a higher level facility, and that higher level facilities are often located in towns where a larger variety of goods and services are available. For these reasons, Oppong [1992] proposes the use of facility location models with probabilistic allocation. Moreover, Rushton [1988] notes that, up to almost a decade ago, models which incorporate consumer choice had not yet been applied to determine the benefits of accessi-

bility improvements in rural areas. These facility location models incorporate aspects of spatial interaction models and a discussion of these models is presented in the next section.

2.3.3 Facility Location Models with Probabilistic Allocation

The previous section reviewed facility location models based on nearest centre allocation. These models have the property that all the demand from a specific user of the system is satisfied by a single new facility, *i.e.*, x_{ij} is limited to being either zero or one. This leads to all users being assigned to the nearest facility. However, as discussed previously, this is not a realistic assumption for flows of users in many real-world systems [O'Kelly, 1987].

Leonardi [1980a] distinguishes two forms of facility location problems: *delivery systems* and *user-attracting systems*. In *delivery systems* users do not travel to obtain services, *i.e.*, the services are delivered from the facilities to the users and the users do not pay for transportation costs. In addition, one decision maker is responsible for both the location of facilities and the allocation of services to users. In this situation, it is reasonable to assume a cost minimization objective and that each user is serviced by the nearest facility [Fotheringham and O'Kelly, 1989]. Thus for a delivery system, a p-median or uncapacitated facility location problem is an appropriate model.

One the other hand, a *user-attracting system* is an appropriate model for many service delivery systems [Leonardi, 1980a]. In this type of system the locational decisions are made by one decision maker, representing a public authority or agency, while the choice of which facility to use is determined by the preferences and choice behaviour of the users. In addition, users travel to the facility and, thus, pay the transport costs while the public agency pays the cost of establishing and maintaining the facility. Clearly, as Leonardi [1980a] notes, a health care system is an example of a user-attracting system with the added complication of exhibiting multiple levels. In these types of systems, the allocation of demand to facilities is stochastic [O'Kelly, 1987]. Thus, in order to model these types of facility location problems, mechanisms must be incorporated to allocate the demand probabilistically to the facilities.

Leonardi [1978] considers the problem of allocating resources to optimize accessibility using an interaction-based approach (as discussed in Section 2.2). His model assumes that the attractiveness of a destination is defined as $f_j^A = (s_j)^{\alpha_A}$ and optimizes

the geometric mean of the Hansen accessibility weighted by the population of each origin zone so that the objective is defined as

$$Z = \prod_i (A_i)^{P_i} \tag{2.59}$$

$$\text{with} \quad A_i = \sum_j s_j^{\alpha_A} D_{ij} \quad . \tag{2.60}$$

By taking the logarithm of the objective function, an equivalent formulation is as follows.

$$\underset{s_j}{\text{Maximize}} \ \sum_i P_i \log A_i \tag{2.61}$$

subject to

$$\sum_j s_j^{\alpha_A} E_j = Q \tag{2.62}$$

$$s_j \geq 0 \qquad j = 1, \ldots, N_D \quad . \tag{2.63}$$

In this model, E_j is the cost of an allocation of unit size in potential location j. Constraint (2.62) ensures that the cost is less that the total budget, Q, while constraints (2.63) ensure that no negative allocations are made. Note that in contrast to other location models, this particular model determines the allocation of resources to various potential facility sites rather than allocating a specified number of facilities. The model is very similar to that proposed by Mayhew and Leonardi [1982] for re-allocating resources to optimize equity and efficiency of their distribution.

Another way of modelling the allocation of a fixed demand to a set of facilities is to use a production-constrained spatial interaction model. Recall that in this model, the flow from origin i to origin j, T_{ij} is

$$T_{ij} = B_i O_i \exp(-\beta D_{ij}) \tag{2.64}$$

$$B_i = 1/\sum_j \exp(-\beta D_{ij}). \tag{2.65}$$

The attractiveness variables in the original formulation, f_j^A, have been removed since they are assumed to be equal. It is easy, however, to generalize the model to incorporate its attraction if this is not the case. Also, the original definition of x_{ij} was the proportion of the flow originating at i that travels to facility j. Thus,

$$T_{ij} = P_i x_{ij}. \tag{2.66}$$

Equations (2.64) and (2.65) are the solutions to the following optimization problem [Fotheringham and O'Kelly, 1989].

$$\underset{x_{ij}}{\text{Minimize}} \ (1/\beta)\sum_{i,j} T_{ij}\log T_{ij} + \sum_{i,j} T_{ij}D_{ij} \tag{2.67}$$

subject to

$$\sum_j T_{ij} = P_i \quad i = 1, \ldots, N_O. \tag{2.68}$$

Thus, the production-constrained model is obtained as the solution to this optimization problem. It should be noted that the objective function (2.64) is the negative of the consumers' surplus [Wilson *et al.*, 1981, p. 171] and thus measures the disbenefit of having to travel to spatially dispersed destinations. This optimization problem is equivalent to that specified for the production-constrained model – namely (2.15), (2.16), and (2.18) – with the total distance constraint being included in the objective as a Lagrangian with a multiplier β.

By adding decision variables to the optimization problem specified by (2.64) and (2.65) and substituting $T_{ij} = P_i x_{ij}$, a p-median type problem can be formulated with the goal of minimizing the users' disbenefit. This problem can be specified as follows.

$$\underset{x_{ij},\, y_j}{\text{Minimize}} \ (1/\beta)\sum_i \sum_{j\in\mathcal{L}} P_i x_{ij}\log\left(b_i x_{ij}\right) + \sum_{i,j} P_i x_{ij} D_{ij} \tag{2.69}$$

subject to the p-median constraints, (2.39) through (2.41), and where $\mathcal{L} = \{j | y_j = 1\}$ is the set of open facilities introduced to avoid calculating $\log 0$. Note that as $\beta \to \infty$ the first term of the objective function drops out and the standard p-median problem remains. It should be noted that Beaumont [1980] derives a similar model for the continuous case based on the maximization of the locational surplus and Leonardi [1980b] derives a series of both attraction-production-constrained models and production-constrained models.

Several different solution methods have been proposed for discrete facility problems with probabilistic allocation. For example, Hodgson [1978] applied a variant of the Interchange method on small attraction-production-constrained test problems. These problems had 10 users and located three new facilities. The Interchange heuristic was found to be robust and relatively efficient for these small test problems. Birkin *et al.* [1995] developed another heuristic method for this problem and applied it to much

larger problems. They found that locating 100 facilities with over 8500 demand nodes and 8000 potential facility locations takes over 150 hours on a Sparc 1 workstation and estimate that locating 1000 facilities would take over 2500 hours.

O'Kelly [1987] developed a method similar to the method proposed by Narula *et al.* [1977] for the p-median problem. The integrality constraints (2.40) were relaxed to be $0 \leq y_j \leq 1$ and a Lagrangian was formed. Similar to the p-median problem, the values of λ_i which maximize the dual problem were found and the values of y_j were then directly calculated. This method was tested on a problem based on data from Hamilton, Ontario. The problem involved locating 20 new facilities among 181 users. He concludes that the efficiency of the algorithm depended on the value of the distance decay parameter, β. The greater the value of β, the larger the gap between the primary and dual objectives. These objectives should be equal at the optimal solution.

Oppong [1992] applied a hierarchical interaction-based facility location model to evaluating the accessibility of primary health care services in Suhum District, Ghana. The model is formulated as follows.

$$\underset{\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3}{\text{Maximize}} \ \sum_k L_k \sum_i P_i \sum_j \exp\left(-\beta_k D_{ij}\right) \tag{2.70}$$

subject to

$$y_{j,k-1} \leq y_{jk} \qquad j = 1,\ldots,N_D \quad k = 2,3 \tag{2.71}$$

$$\sum_j y_{jk} = \sum_{\ell=1}^{k} p_\ell \qquad k = 1,2,3 \tag{2.72}$$

$$y_{jk} \in \{0,1\} \qquad j = 1,\ldots,N_D \quad k = 1,2,3 \ . \tag{2.73}$$

In this model, L_k is the level-specific attractiveness for a facility of a given order and β_k is the level-specific distance decay parameter. This model attempts to maximize the overall aggregate benefit from the configuration of the system. The values of L_k and β_k were calibrated from fitting a spatial interaction model using actual utilization data from the district. Oppong concludes that the lowest level of facilities had very little attractiveness and that it is important to ensure that "available health facilities provide a certain minimum level of service that is acceptable to users" (p. 170). However, he notes that the results of this model must be applied with caution since they may exacerbate existing urban/rural disparities. This may be due to the biased sample (those who attend health facilities) used to calibrate the model. Oppong states that the need for research to resolve such difficulties is critical.

2.3.4 Multicriteria Location Problems

The problem of facility location planning and decision-making for primary health care provision is typically a complex spatial problem. Here, it is important not only to plan the system to be efficient but, for reasons noted in Section 2.1, it is also vital that the distribution of resources is equitable and that these resources are universally accessible [WHO, 1994]. For example, Oppong [1992] points out that results of his interaction-based model tend to increase rural/urban disparity. One reason for this is that he used an efficiency objective which maximized the aggregate level of benefit and thus concentrated resources in areas with the most population. If he had used an equity objective which aims to equalize the distribution of resources, this result would not have occurred. In addition to maximum distance and coverage measures, Mulligan [1991] notes several possible measures of equity in facility location models such as mean deviation, concentration indices, Gini coefficients and variance.

From this discussion, it is evident that location planning for primary health care provision is a multicriteria location problem. Such a problem can be structured as follows. Define $Z_1, Z_2, \ldots, Z_{N_C}$ to be the N_C objective functions.

$$\text{Minimize } \mathbf{Z}(\mathbf{z}) \tag{2.74}$$

$$\text{subject to } \mathbf{z} \in \mathcal{A} \tag{2.75}$$

where $\mathbf{Z} = (Z_1, \ldots, Z_{N_C})$ represents a vector of N_C criteria. A solution $\mathbf{Z}'$ is dominated by another solution $\mathbf{Z}''$ if

$$\mathbf{Z}' \neq \mathbf{Z}'' \text{ and } Z'_k \leq Z''_k \text{ for all } k = 1, \ldots, N_C.$$

However, in the case of conflicting objectives, there often does not exist a single solution that dominates all other solutions. A locational alternative or solution is termed efficient (or non-dominated) if it is feasible and no other feasible locational alternative can improve on one criterion without reducing the performance of another.

Malczewski and Ogryczak [1995] outline two main techniques for generating efficient solutions of multicriteria location problems. These include the constraint method and the weighting method. The constraint method involves optimizing one objective, Z_ℓ and setting maximum allowable levels for the other criteria, ε_k. Thus, the multicri-

teria problem is transformed into a single-criterion problem:

$$\text{Minimize } Z_\ell(\mathbf{z}) \tag{2.76}$$

$$\text{subject to } \mathbf{z} \in \mathcal{A} \tag{2.77}$$

$$Z_k(\mathbf{z}) \leq \varepsilon_k, \quad \text{for } k = 1, \ldots, N_C, k \neq \ell \quad . \tag{2.78}$$

The set of efficient solutions to the problem can be generated by parametric variation of the ε_k [Malczewski and Ogryczak, 1995].

The weighting method involves assigning a weight, $\omega_k \geq 0$, to each of the objective functions and solving the single-criterion problem:

$$\text{Minimize } \sum_k \omega_k Z_k(\mathbf{z}) \tag{2.79}$$

$$\text{subject to } \mathbf{z} \in \mathcal{A}. \tag{2.80}$$

The set of efficient solutions can be found through parametric variation of the weights [Malczewski and Ogryczak, 1995].

An example of using a multicriteria location problem for health care planning in developing countries is provided by Massam and Malczewski [1991]. In this study, they find the best site for a health centre in rural Zambia. Six criteria are defined for evaluating the decisions: average weighted distance, standard deviation of the distance, maximum distance, population within 12 km, population within 30 km, and distance to the nearest centre. Each of the objectives was optimized separately to calculate the best values and the worst values for each objective. Finally, various aspiration levels were calculated for the objectives and suitable alternatives were selected depending on the importance that decision makers placed on the different criteria.

2.3.5 Summary of Facility Location Models

This section discussed aspects of discrete or site-selecting facility location problems. Several different types of such problems were outlined and their respective solution methods were discussed.

First, an overview of facility location models for the delivery system was provided. These models assume that all the users at a given location attend the nearest facility. The p-median problem involved locating p facilities so that the total weighted distance from

the demand nodes to the facilities is minimized. The uncapacitated facility location problems included a term for the cost of establishing a facility and minimized the total cost. The set covering problem determined a set of facilities so that each demand node was within a maximum distance of a facility while the maximum covering problem attempted to locate p facilities to maximize the demand that was within a maximum distance standard. The p-centre problem located facilities so that the maximum distance from any demand node to a facility is minimized. Finally, the hierarchical location problem requires the facilities to be organized into a hierarchy. Several examples of these problems applied to health system planning were then discussed.

Next, a facility location model for user-attracting systems was discussed. In these systems the users choose the facility they attend according to some distance decay effect. This model combines spatial interaction models with facility location problems. First, a simple model based on accessibility maximizing was presented. Next, a mathematical formulation for locating p facilities was introduced for this type of system combining an origin-constrained spatial interaction model and a p-median problem. A solution method for this model was briefly discussed. A hierarchical model based on maximizing aggregated benefits was also discussed. Finally, it was noted that facility location models for primary health care optimization in developing countries really involve the examination of several different objectives and, thus, multiobjective problem formulations are appropriate.

2.4 Chapter Summary

This chapter reviewed a selection of existing studies, organized within a conceptual framework, encompassing health care accessibility. Further, spatial interaction models and facility location models were discussed, and examples were presented of the application of these two types of mathematical models to develop strategies for improving accessibility to health care services.

As noted previously in this chapter, accessibility to health care services is a "slippery" concept to define [Phillips, 1990]. Further, many different measures of potential accessibility have been proposed. These measures, although intuitively reasonable, often do not provide an obvious justification for their specific mathematical form. The next chapter expands upon Khan and Bhardwaj's [1994] typology of accessibility by

providing a generic mathematical framework in which to consider potential accessibility measures.

Chapter 3

Accessibility to Health Care

This chapter introduces a generic model of potential accessibility to health care that allows the concepts discussed in the previous chapter to be transformed meaningfully into objective and measurable terms. The focus is on how accessibility can be measured, and more specifically, on how the distribution and characteristics of supply points or facilities affect accessibility.

A generic model for potential accessibility to primary health care is discussed and presented. Although the generic model outlines the properties of a potential accessibility measure, it does not provide a specific interpretation for the measure. A behavioural framework based on an individual choosing the alternative with the highest level of attractiveness is then discussed in order to provide a rationale for specific accessibility measures. This framework is used to develop several measures of potential accessibility.

3.1 A Generic Model for Potential Accessibility Measures

As mentioned in the previous chapter, it is important to differentiate between realized and potential accessibility to primary health care. Realized accessibility relates to actual health utilization patterns. However, several researchers have noted weaknesses with realized accessibility in that it is very difficult to define "need" for health care, and also, there is a multiplicity of factors intervening between the concept of need and the use of available services [Bradshaw, 1972; Fielder, 1981]. In contrast, potential accessibility is much more narrowly defined. It emphasizes the opportunity or potential for individual

behaviour rather than actual behaviour. The data requirements for accurate estimation of potential geographic accessibility are much less stringent than for realized accessibility. This is a particularly important consideration for health facility location planning in developing countries. Towards this end, this section outlines a generic model for measuring potential access to health care services.

Geographic, or spatial, accessibility refers to the level of difficulty an individual has in obtaining services from a service provider. If a sub-area is inaccessible, it is, for all intents and purposes, very difficult for an individual living there to obtain services; while in an accessible region the consumption of services is relatively easy. Thus, accessibility relates to the ease of spatial interaction, or the potential opportunity for spatial interaction between the service supply nodes and the target users or consumers (demand) [Weibull, 1980]. An accessibility measure in this generic model converts this potential for interaction over space into a non-negative real number.

3.1.1 Definitions

Using the concepts of service supply, demand, and spatial interaction, it is possible to establish a framework that defines the accessibility properties of a particular system. The two elements interacting within this system are the potential users and the facilities or service providers. Furthermore, it is possible to define a spatial separation between each user and each facility.

Users: The system has a total population of N_U potential users. Each potential user i has an associated vector of N_P characteristics, $\mathbf{X}_i^U = \left(X_{i1}^U, X_{i2}^U, \ldots, X_{i,N_P}^U\right)$. Let $\mathcal{U} = \{\mathbf{X}_1^U, \mathbf{X}_2^U, \ldots, \mathbf{X}_{N_U}^U\}$ represent the set of potential users for the system and let $\mathcal{N}_U = \{1, 2, \ldots, N_U\}$ be the corresponding index set.

Facilities: The relevant services for the system are provided by N_D facilities or service providers. Each facility has N_A relevant attributes represented by the vector $\mathbf{X}_j^F = \left(X_{j1}^F, X_{j2}^F, \ldots, X_{j,N_A}^F\right)$. Again, define the set $\mathcal{F} = \{\mathbf{X}_1^F, \mathbf{X}_2^F, \ldots, \mathbf{X}_{N_A}^F\}$ as the set of facilities and the corresponding index set $\mathcal{N}_D = \{1, 2, \ldots, N_D\}$.

Spatial Separation: Between each potential user and each facility there is a vector of spatial separation values. Define the function $\mathbf{S}$ that maps a given potential

user/facility pair onto a vector of N_S non-negative real values[1]

$$\mathbf{S} : (\mathcal{N}_{\mathcal{U}} \times \mathcal{N}_{\mathcal{D}}) \rightarrow \mathbb{R}_{+}^{N_S}$$

with set $S = \{\mathbf{S}(i,j) | (i,j) \in \mathcal{N}_{\mathcal{U}} \times \mathcal{N}_{\mathcal{D}}\}$.

From these definitions it is possible to define the properties of a generic accessibility measure.

DEFINITION 1 An accessibility measure for a given individual $i \in \mathcal{N}_{\mathcal{U}}$ is a function that maps from a given set of users and facilities onto a finite non-negative real number

$$A_i = f_i(\mathcal{U}, \mathcal{F}) \rightarrow \mathbb{R}_{+}.$$

Moreover, this function has the following three properties:

$$f_i(\mathcal{U}, \emptyset) = 0 \qquad \text{(I)}$$

$$f_i(\mathcal{U}, \mathcal{F}') \leq f_i(\mathcal{U}, \mathcal{F}) \qquad \text{for } \mathcal{F}' \subset \mathcal{F} \qquad \text{(II)}$$

$$f_i(\mathcal{U}', \mathcal{F}) \leq f_i(\mathcal{U}, \mathcal{F}) \qquad \text{for } \mathcal{U} \subset \mathcal{U}' \qquad \text{(III)}$$

for all $i \in \mathcal{N}_{\mathcal{U}}$.

A system is considered completely *inaccessible* to a particular user if that user's accessibility is zero, *i.e.*, $A_i = 0$. The three properties establish reasonable behaviour for an accessibility measure. The first property states that a system with no facilities is inaccessible. The second property ensures that accessibility cannot decrease with an increasing number of facilities. The last property states that accessibility cannot increase with an increasing number of potential users. Note that properties (I) and (II) imply that the range of the accessibility function is the positive real numbers.

The generic model assumes that the accessibility measure is *aggregable* so that it is possible to define the accessibility level of an individual to each facility. The overall accessibility of an individual *to the system* can be expressed as a function of the facility

[1]The $\times$ symbol in this discussion is used to denote the Cartesian product of two sets. If $\mathcal{A}$ and $\mathcal{B}$ are sets, then the Cartesian product of $\mathcal{A}$ and $\mathcal{B}$ consists of the set of all ordered pairs having the first element in $\mathcal{A}$ and the second element in $\mathcal{B}$. Thus $\mathcal{A} \times \mathcal{B} = \{(a,b) | a \in \mathcal{A} \text{ and } b \in \mathcal{B}\}$.

accessibility values for the individual. This property is restated formally in section 3.1.4. Three factors characterizing the accessibility of a facility are now developed.

3.1.2 Facility-Dependent Factors

For a given individual $i \in \mathcal{N}_{\mathcal{I}}$, it is proposed that the accessibility to a facility $j \in \mathcal{N}_{\mathcal{D}}$ is affected by three facility-dependent factors. The first two factors, attraction, a_{ij}, and distance, d_{ij}, are standard factors affecting accessibility (cf. Weibull [1976; 1980]). However, both Joseph and Bantock [1982] and Leonardi [1980a] note the importance of congestion effects. Therefore, the accessibility framework is modified to allow for this third effect by incorporating a congestion factor, c_{ij}.

Attraction: Attraction is a facilitator for accessibility, although specific attributes of the facility may affect attraction either positively or negatively. Factors such as the quality of services, the size of facility, the level of supply, and the type and mix of services offered can affect the level of attraction. Thus, the attraction of a facility captures information on how organizational characteristics of the facility influence an individual's accessibility. The characteristics that influence an individual's attraction to a facility are location-independent and dependent only on that facility's attributes so that for an individual i and a facility j

$$a_{ij} = G_A\left(\mathbf{X}_j^F\right) \tag{3.1}$$

where G_A is the facility attraction function.

Distance: An important barrier to accessibility is the distance between the individual and the facilities. The distance is a scalar non-negative real number representing the difficulty that an individual i has in reaching facility j and is a function of N_S spatial separation variables so that

$$d_{ij} = G_D\left(\mathbf{S}(i,j)\right) \tag{3.2}$$

where G_D is the distance function. In most measures, there is only a single distance factor so that $d_{ij} = \mathbf{S}(i,j) = D_{ij}$. However, it is possible that several factors, related to the location of a user with respect to a facility, may influence distance.

Congestion: Congestion can also be an important factor affecting accessibility. An overcrowded facility can result in long waiting times that can act as barriers to use and reduce the effective level of accessibility. For example, Rosero-Bixby [1995] states that in Costa Rica the median reported travel time to public family planning outlets is 28 minutes while the median waiting time is 141 minutes. Thus, congestion can play an important role in determining accessibility, particularly in areas with a large population and few resources. Although congestion may be included as a negative influence on attraction, the generic model, defined here, separates these factors, due to their different causes. Attraction is affected by facility-specific factors while congestion is dependent upon the interaction between a facility and its surrounding population. Thus, the congestion for a given facility is related to the attributes of the facility and its relative location with respect to potential users and can be expressed for individual i and facility j as

$$c_{ij} = G_C\left(\mathbf{x}_j^F, \mathcal{U}, \mathcal{S}\right) \tag{3.3}$$

where G_C is the facility congestion function. The congestion of a facility cannot decrease with an increase in the number of potential users. This implies the following property of the congestion function when $\mathcal{U} \subset \mathcal{U}'$

$$G_C\left(\mathbf{x}_j^D, \mathcal{U}, \mathcal{S}\right) \leq G_C\left(\mathbf{x}_j^D, \mathcal{U}', \mathcal{S}\right). \tag{3.4}$$

3.1.3 Facility Accessibility Function

The three facility-dependent factors, attraction, distance, and congestion, affect the accessibility of a particular facility for a given individual. Therefore, the accessibility of individual i to facility j is given by

$$A_{ij} = g_{ij}\left(a_{ij}, d_{ij}, c_{ij}\right) \rightarrow \mathbb{R}_+ \tag{3.5}$$

where $g_{ij}(\cdot)$ is the corresponding facility accessibility (FA) function, and a_{ij}, d_{ij}, and c_{ij} are defined by equations (3.1), (3.2), and (3.3) respectively. A facility is considered *inaccessible* to a particular user if the corresponding FA function value is zero, *i.e.*, $A_{ij} = 0$. Different accessibility measures are derived from different facility accessibility functions. However, it is possible to establish some conditions on the FA function because

attraction facilitates accessibility while congestion and distance act as barriers.

It is important to note that the characteristics of the individual, X_i^U, influence how the attraction, distance, and congestion affect an individual's accessibility to a facility. For example, the type of transportation available to an individual would modify how distance affects accessibility. Similarly, the presence of organizational barriers at a facility for an individual would lower the accessibility of that facility.

A facility that has zero attraction (due to, for example, a complete lack of resources) is considered to be inaccessible to all individuals, that is, if $a_{ij} = 0$ for some $j \in \mathcal{F}$, then

$$A_{ij} = g_{ij}\left(a_{ij}, d_{ij}, c_{ij}\right) = 0$$

for all individuals $i \in \mathcal{U}$. Further, the level of accessibility should not decrease as attraction increases. Mathematically, these conditions can be expressed as follows

$$g_{ij}(a,d,c)\big|_{a=0} = 0 \tag{3.6}$$

$$g_{ij}(a,d,c) \geq g_{ij}\left(a',d,c\right) \quad \text{for } a' > a \tag{3.7}$$

for all $i \in \mathcal{U}$ and $j \in \mathcal{F}$.

Similar conditions for the effect of distance on accessibility can be found, however, these effects are more complicated [Joseph and Phillips, 1984]. For instance, Girt [1973] found that distance can have both a positive and a negative effect on utilization behaviour. Moreover, social and psychological influences may have an effect on how distance influences accessibility, particularly for family planning services. A family planning clinic located in the same community that a woman lives in may be effectively inaccessible due to social stigma associated with, for example, contraceptive usage. The woman may instead prefer the anonymity of patronizing a more distant facility in a different community. Therefore, the conditions relating distance to accessibility are modified. The model makes no assumptions about the effect of distance within the restricted range R', but beyond this range, accessibility cannot increase with increasing distance and the facility becomes inaccessible when it is very distant. These relations can be expressed as follows

$$\lim_{d\to\infty} g_{ij}(a,d,c) = 0 \tag{3.8}$$

$$g_{ij}(a,d,c) \leq g_{ij}(a,d',c) \quad \text{for } d' \geq d \geq R' \geq 0. \tag{3.9}$$

Typically, there is a maximum service range $R \geq R'$ beyond which a facility is considered inaccessible. With this assumption, condition (3.8) becomes

$$g_{ij}(a,d,c)\,|_{d \geq R} = 0. \tag{3.10}$$

Conditions can also be established on how congestion affects accessibility. A facility with a very high level of congestion should be considered inaccessible and the accessibility cannot increase with increasing congestion. The mathematical expressions for these conditions are as follows.

$$\lim_{c \to \infty} g_{ij}(a,d,c) = 0 \tag{3.11}$$

$$g_{ij}(a,d,c') \leq g_{ij}(a,d,c) \quad \text{for } c' > c. \tag{3.12}$$

The next two subsections discuss how to aggregate these facility accessibility values into an overall system accessibility measure.

3.1.4 Aggregable Accessibility Measures

As mentioned previously, the generic model assumes that the accessibility measure is *aggregable*. An accessibility measure is termed aggregable if there exists a representation of the accessibility measure such that

$$A_i = f_i(\mathcal{U}, \mathcal{F}) = \oplus\left(A_{i1}, \ldots, A_{i,N_D}\right) \tag{3.13}$$

for all $i \in N_U$, where A_{ij} is defined as in equation (3.5) and $\oplus$ is the aggregation operator. An alternative functional form for this expression is

$$A_i = A\left(\mathbf{X}_i^P, \mathcal{F}, \mathbf{D}_i\right) \tag{3.14}$$

where $\mathbf{D}_i$ is the vector of distances to the facilities[2].

For notational convenience, in the following descriptions it is assumed that the accessibility is being evaluated for a given individual. Therefore, the index i is unneces-

[2]This can be shown through the appropriate substitutions and the assumption that the facility congestion values are absorbed into the vector of facility-specific characteristics.

sary and omitted, and the number of facilities is defined as $N = N_D$. The aggregation operator has the following four properties.

Commutativity: The aggregation operator is commutative, *i.e.*, the order in which the facilities are listed does not affect the value:

$$\oplus(A_1, \ldots, A_N) = \oplus\left(A_{j_1}, \ldots, A_{j_N}\right) \tag{3.15}$$

where $(j_1, \ldots, j_N) = \sigma(1, \ldots, N)$ and σ represents a permutation operation.

Monotonicity: The overall accessibility does not decrease with any increase in the facility accessibility levels. This means that $A'_j > A_j$ implies

$$\oplus\left(A_1, \ldots, A'_j, \ldots, A_N\right) \geq \oplus\left(A_1, \ldots, A_j, \ldots, A_N\right). \tag{3.16}$$

Zero Identity Element: The addition of an inaccessible facility does not affect the overall accessibility. This implies that zero is the identity element for the aggregation operator

$$\oplus(A_1, \ldots, A_{N-1}, 0) = \oplus(A_1, \ldots, A_{N-1}). \tag{3.17}$$

Since $\oplus$ is commutative, there is no change in the overall level of accessibility through the addition or removal of any number of inaccessible facilities.

Non-Negativity: The accessibility of a system consisting of a single inaccessible facility is zero

$$\oplus(0) = 0 \quad . \tag{3.18}$$

Since $\oplus$ is monotonic, this implies that accessibility must be non-negative, since $A_j \geq 0$ by definition.

An aggregable accessibility measure meets all the conditions of an accessibility measure as specified in DEFINITION 1. To prove this proposition, properties (I), (II), and (III) are now proved in order. The first property holds because of non-negativity and because $\oplus$ has a zero identity element.

PROOF 1 *To prove property* (I), *requires proving that* $f(\mathcal{U}, \emptyset) = 0$.

$$\begin{aligned} f(\mathcal{U}, \emptyset) &= \oplus(\,) && \text{by definition} \\ &= \oplus(0) && \text{by (3.17)} \\ &= 0 && \text{by (3.18)} \quad \diamond \end{aligned}$$

The second proof involves showing that system accessibility is a non-decreasing function of the total number of facilities.

PROOF 2 *Consider a set* $\mathcal{F}' \subset \mathcal{F}$ *with* $N' < N$ *elements. It is required to prove that property* (II) *holds, namely that*

$$f(\mathcal{U}, \mathcal{F}) \geq f(\mathcal{U}, \mathcal{F}').$$

Define $\mathcal{N} = \{1, \ldots, N\}$ *and* $\mathcal{N}' = \{1, \ldots, N'\}$. *Let* A_j *be the accessibility of facility* $j \in \mathcal{N}$ *and define* A'_j *as the accessibility of facility* $j \in \mathcal{N}'$. *Furthermore, construct* A''_j *so that*

$$A''_j = \begin{cases} A_j & \text{if facility } j \text{ is in set } \mathcal{F}', \\ 0 & \text{otherwise,} \end{cases}$$

for $j \in \mathcal{N}$, *and the permutation* $(j_1, \ldots, j_N) = \sigma(1, \ldots, N)$ *such that*

$$A''_{j_k} = \begin{cases} A'_k & \text{if } k \in \mathcal{N}', \\ 0 & \text{otherwise.} \end{cases}$$

It follows that

$$\begin{aligned} f(\mathcal{U},\mathcal{F}) &= \oplus(A_1,\ldots,A_N) && \text{by definition} \\ &\geq \oplus(A''_1,\ldots,A''_N) && \text{by (3.16)} \\ &= \oplus\left(A''_{j_1},\ldots,A''_{j_N}\right) && \text{by (3.15)} \\ &= \oplus(A'_1,\ldots,A'_{N'},0,\ldots,0) && \text{by construction} \\ &= \oplus(A'_1,\ldots,A'_{N'}) && \text{by (3.17)} \\ &= f(\mathcal{U},\mathcal{F}') && \diamond \end{aligned}$$

The final proof involves showing that accessibility cannot increase for an increasing number of users.

PROOF 3 *For a set $\mathcal{U} \subset \mathcal{U}'$ with $N_U \leq N'_U$ elements, to prove property* (III) *requires a proof that*

$$f(\mathcal{U}',\mathcal{F}) \leq f(\mathcal{U},\mathcal{F})$$

for any user $i \in \mathcal{U}$.

Define $\mathcal{N}_{\mathcal{U}} = \{1,\ldots,N_U\}$, $\mathcal{N}_{\mathcal{U}}{}' = \{1,\ldots,N'_U\}$, and set $\mathcal{N}_{\mathcal{D}} = \{1,\ldots,N_D\}$ along with corresponding definitions of $\mathcal{S}$ and $\mathcal{S}'$. Further, for a given user $i \in \mathcal{N}_{\mathcal{U}}$, define

$$c_j = G_C\left(X_j^D,\mathcal{U},\mathcal{S}\right) \quad \text{and} \quad c'_j = G_C\left(X_j^D,\mathcal{U}',\mathcal{S}'\right)$$

and

$$A_j = g_j(a_j,d_j,c_j) \quad \text{and} \quad A'_j = g_j\left(a_j,d_j,c'_j\right)$$

by definitions (3.3) *and* (3.5) *respectively where a_j and d_j are the attraction and distance of facility j.*

Thus

$$
\begin{aligned}
& & f(\mathcal{U},\mathcal{F}) &= \oplus\left(A_1,\ldots,A_{N_D}\right) && \textit{by definition}\\
\textit{But,} & & c_j &\leq c'_j && \textit{by (3.4)}\\
\textit{Consequently,} & & A_j &\geq A'_j && \textit{by (3.12)}\\
\textit{Therefore,} & \oplus\left(A_1,\ldots,A_{N_D}\right) &&\geq \oplus\left(A'_1,\ldots,A'_{N_D}\right) && \textit{by (3.16)}\\
& & &= f(\mathcal{U},\mathcal{F}') && \diamond
\end{aligned}
$$

3.1.5 Separable Accessibility Measures

An important category of aggregable accessibility measures are *separable* accessibility measures. In fact, Weibull [1980] notes that most accessibility measures currently in use are separable. A separable measure is an aggregable accessibility measure in which one facility does not affect the accessibility of another facility, *i.e.*, the facilities are independent of each other with respect to accessibility. More specifically, we can distinguish two types of separable measures: *strictly separable measures* and *transform-separable measures*.

An accessibility measure is termed strictly separable if the aggregation operator is binary and associative. Thus, a strictly separable accessibility measure can be represented as

$$A_i = A_{i1} \oplus A_{i2} \oplus \cdots \oplus A_{i,N_D}. \tag{3.19}$$

In this equation $\oplus$ is a *binary aggregation operator* with the properties (3.15), (3.16), (3.17), and (3.18) and is associative, *i.e.*,

$$\oplus\left(A_i, A_j, A_k\right) = (A_i \oplus A_j) \oplus A_k = A_i \oplus (A_j \oplus A_k). \tag{3.20}$$

An alternative way of expressing associativity is as follows [Fodor and Roubens, 1994]:

$$\oplus^{(N)}(A_1,\ldots,A_n) = \oplus^{(2)}\left(\oplus^{(n-1)}(A_1,\ldots,A_{n-1}), A_n\right) \tag{3.21}$$

where $\oplus^{(N)}$ represents aggregation over N accessibilities. For a separable accessibility measure, the aggregation operator is defined in terms of combining two accessibilities. Through the associative property, this operator can be canonically extended to any finite

number of facilities. This property of associativity implies that, in a strictly separable accessibility measure, there is no interaction between facilities. Each facility contributes independently to the overall system accessibility for a given individual.

A transform-separable accessibility measure is a relaxation of a strictly separable measure. For a transform-separable measure, the accessibility is some strictly increasing monotonic function of a strictly separable accessibility measure, *i.e.*,

$$A_i = T\left(A_{i1} \oplus A_{i2} \oplus \cdots \oplus A_{i,N_D}\right) \tag{3.22}$$

where T is the transformation function. In other words, a function is transform-separable if there exists a function T^{-1} such that $T^{-1}(A_i)$ is strictly separable. In order to illustrate the difference between a strictly separable and a transform-separable accessibility measure, consider the following simple example.

EXAMPLE 1 *Let $\oplus$ be the aggregation operator for an accessibility measure where $\oplus$ is defined as*

$$A = \oplus^{(N)}(A_1, A_2, \ldots, A_n) = \ln\left(1 + \sum_{j=1}^{n} A_j\right).$$

Trivially, this operator is commutative and has zero as the identity element. Moreover, $\frac{\partial A}{\partial A_k} = 1/(1 + \sum_{j=1}^{n} A_j)$ such that it is also monotonically increasing, but it is not associative. Consider a system with the facilities having accessibilities of A_1, A_2, and A_3. The accessibility of this system is

$$A = \oplus^{(3)}(A_1, A_2, A_3) = \ln(1 + A_1 + A_2 + A_3)$$

but,

$$\oplus^{(2)}\left(\oplus^{(2)}(A_1, A_2), A_3\right) = \ln[1 + \ln(1 + A_1 + A_2) + A_3] \neq A$$
$$\oplus^{(2)}\left(A_1, \oplus^{(2)}(A_2, A_3)\right) = \ln[1 + A_1 + \ln(1 + A_2 + A_3)] \neq A \quad .$$

Therefore, since the aggregation operator is commutative and monotonic with zero as the identity element, the accessibility measure defined by this operator is an aggregable measure. But the operator is not associative so that this measure is not a separable

measure. However, if we define $T^{-1}(x) = \exp x - 1$ *and note that*

$$T^{-1}\left[\oplus^{(N)}(A_1, A_2, \ldots, A_n)\right] = \sum_{j=1}^{n} A_j$$

then, since the addition operator is commutative, monotonic, and associative with zero as an identity element, this accessibility measure is a transform-separable measure.

Weibull [1980] identifies two important types of separable accessibility measures: *additive* measures and *maxitive* measures. A separable accessibility measure is said to be an *additive accessibility measure* if the aggregation operator is defined as

$$\oplus^{(N)}(A_1, A_2, \ldots, A_N) = \sum_j A_j. \tag{3.23}$$

This operator meets the conditions of commutativity, monotonicity, and associativity, has zero as the identity element, and is consistent with $\oplus(0) = 0$. Traditional gravity model-based accessibility measures are additive.

Maxitive accessibility measures define the aggregation operator to be

$$\oplus^{(N)}(A_1, A_2, \ldots, A_N) = \max_j A_j. \tag{3.24}$$

Again, the maximum operator meets all the conditions of a separable measure. Suitably transformed, the minimum distance measure can be considered a maxitive accessibility measure. Weibull [1980] suggests that maxitive measures result from assuming a choice process where the individual selects the facility offering the maximum attractiveness among all facilities.

In summary, accessibility measures capture the potential level of interaction between an individual and a system providing services. The model proposes that the accessibility of an individual to a facility is a function of the facility's attraction, distance, and congestion. Accessibility measures are assumed to be aggregable so that they can be combined into an overall system-wide accessibility value for an individual using an aggregation operator. Separable accessibility measures are an important class of accessibility measures so that each facility is considered independently. Thus, this generic model provides a flexible model for describing accessibility measures. Nevertheless, the model lacks behavioural interpretation. The next section develops such an interpretation for the model.

3.2 An Attractiveness Maximization Framework for the Generic Accessibility Model

The form of a specific accessibility measure depends on individual behaviour. For example, if an individual always chooses the nearest facility then the only factor that affects accessibility would be distance to the nearest facility. On the other hand, if the users do not behave in this manner, and there is considerable evidence that generally they do not [Joseph and Phillips, 1984; Martin and Williams, 1992], then alternative measures would be more appropriate. Indeed, any measure of system-wide accessibility must be considered as the result of an individual decision process. Thus, the individual decision-making process can be considered fundamental to measuring accessibility. This section develops a behavioural interpretation for the generic model formulated in Section 3.1 in terms of individual choice theory.

Consider a system, as defined in the previous section, consisting of a set of spatially distributed service providers or facilities, $\mathcal{F}$, and a set of spatially distributed clients or users in the target population group, $\mathcal{U}$. The choice problem examines the question of which particular service provider, if any, a particular individual chooses. Ben-Akiva and Lerman [1985] define a choice problem in terms of four elements: the decision makers, the alternatives, the attributes of alternatives, and the decision rule.

The Decision Makers: The decision makers in a choice problem are the individuals in the target population group. These individuals face different choice situations and have differing needs, desires, and tastes. Thus, important factors in any choice situation are the characteristics of the individual making the choice.

The Alternatives: Any choice is made from a non-empty set of alternatives. The total range of potential alternatives is deemed the *universal set* of alternatives while the set of feasible alternatives for a given individual is that individual's *choice set*. In our context, the universal set of alternatives is based on the set of facilities augmented by the "null" option, *i.e.*, an alternative for the decision not to choose any facility.

Alternative Attributes: Within the choice problem is the assumption that each alternative can be characterized in terms of its attributes. There are four main sources of these attributes in an accessibility context. The attributes of the individual capture the needs and tastes of the individual. The attributes of a facility determine

its attraction. In addition, each facility has a vector of spatial separation attributes for a given individual. And finally, facilities have associated attributes related to congestion.

The Decision Rule: The final element of a choice model is the decision rule. A decision rule describes the internal strategy that the decision maker uses to process the available information and select a particular alternative. Both Ben-Akiva and Lerrman [1985] and Fotheringhman and O'Kelley [1989] describe different decision rules and information processing strategies. The accessibility model assumes commensurability of the attributes so that it is possible to reduce the vector of attributes to a scalar value. This value is termed the *attractiveness*[3] of the alternative and establishes a preference ordering among the alternatives. If one alternative has a higher attractiveness than another then the individual would prefer that alternative. This decision rule is that the individual selects the alternative with the maximum attractiveness.

In order to express this problem mathematically, some definitions are required. Consistent with the earlier definitions, define the index set of potential users as $\mathcal{N}_{\mathcal{U}}$, and let the index set of facilities be given by $\mathcal{N}_{\mathcal{D}}$. Thus the universal set of alternatives is $\mathcal{A} = \mathcal{N}_{\mathcal{D}} \cup \{0\}$ where 0 is the null alternative and the feasible choice set for a given individual is $\mathcal{A}_i \subseteq \mathcal{A}$, $i \in \mathcal{N}_{\mathcal{U}}$. Let the client attributes associated with individual $i \in \mathcal{N}_{\mathcal{U}}$ be denoted by $\mathbf{X}_i^P$, the attributes of facility $j \in \mathcal{N}_{\mathcal{D}}$ by $\mathbf{X}_j^F$ which include congestion attributes, and the separation attributes between an individual i and a facility j be denoted by $\mathbf{X}_{ij}^D$, $(i,j) \in \mathcal{N}_{\mathcal{U}} \times \mathcal{N}_{\mathcal{D}}$. It is notationally convenient to define a new vector of attributes $\mathbf{Y}$ that combines $\mathbf{X}_i^P$, $\mathbf{X}_j^F$, and $\mathbf{X}_{ij}^D$ for a given individual-alternative pair so that

$$\mathbf{Y}_{ij} = \mathbf{h}\left(\mathbf{X}_i^P, \mathbf{X}_j^F, \mathbf{X}_{ij}^D\right) \tag{3.25}$$

where $\mathbf{h}$ is a vector-valued function. The choice problem can then be stated as follows:

> *Given an individual i from the set of clients for the system, $i \in \mathcal{N}_{\mathcal{U}}$, which, if any, facility does this particular individual choose from the set of feasible service providers, $j \in \mathcal{N}_{\mathcal{D}}$, available to that individual.*

[3]The term "attractiveness" is used here instead of the standard term utility in order to emphasize that this attractiveness need not meet any specific properties of a utility measure.

Thus, define the attractiveness

$$U_{ij} = U\left(\mathbf{Y}_{ij}\right) \tag{3.26}$$

where U_{ij} is the attractiveness of alternative j to individual i and $U(\cdot)$ is the perceived attractiveness function, which maps from the vector of attributes onto the set of real numbers. Note that U_{i0} is the attractiveness of the null alternative, *i.e.*, non-attendance.

In the context of these choice models, the accessibility measure is a scalar summary of the difference in the "satisfaction" of an individual between the presented system and the null alternative. A reasonable value for the satisfaction is the attractiveness of the selected alternative [Daganzo, 1979]. Thus, given a choice set $\mathcal{A}_i$ for an individual i, the accessibility is some monotonic function of the selected alternative – the alternative with the maximum attractiveness, *i.e.*,

$$U_i^* = \max_{j \in \mathcal{A}_i} U_{ij} \tag{3.27}$$

$$A_i = T\left(U_i^* - U_{i0}\right) = T\left(\max_{j \in \mathcal{A}_i} U_{ij} - U_{i0}\right) \tag{3.28}$$

where T is a function transforming satisfaction into accessibility. Similarly, the accessibility of facility j to individual i, A_{ij}, is the difference in the satisfaction between a system consisting of that facility and an empty system,

$$A_{ij} = T\left[\max\left(U_{ij}, U_{i0}\right) - U_{i0}\right]. \tag{3.29}$$

A choice situation with a higher satisfaction should correspond to a higher level of accessibility. Therefore, T is a monotonic function of the satisfaction. Most existing accessibility measures can be defined using either a linear function T_1 or an exponential function T_2 where

$$T_1(U) = \omega U \qquad \text{or} \qquad T_2(U) = \exp(\omega U) \tag{3.30}$$

with $\omega > 0$ a constant.

In addition, for each individual i, the probability of selecting alternative j is given by

$$p_{ij} = \Pr\left[U_{ij} \geq U_{ik}, \forall k \in \mathcal{A}_i\right] \tag{3.31}$$

and can be interpreted as the probability that a given alternative's attractiveness is greater than any other alternative. Assuming that no two alternatives have the same level of attractiveness, then $p_{ij} = 1$ for the alternative with the highest attractiveness and 0 for all other alternatives.

Ben-Akiva and Lerman [1985] identify two important issues in specifying the attractiveness functions: the attributes and the functional form. Fotheringham and O'Kelly [1989, p. 70] note that there is "a great diversity of attributes that appear to be relevant across different choice situations." Moreover, measurements of the actual attributes often depend on the availability of data [Daganzo, 1979]. Three broad influences were identified in the generic model introduced in the previous section (attraction, distance, and congestion).

The second important issue is determining the functional form of U_{ij}. It is important that this function accurately reflect how the various attributes affect attractiveness. Moreover, it is also important that the function has convenient computation properties for estimating the values of unknown parameters. One widely-used functional form for the attractiveness function is a *linear-in-parameters* form [Ben-Akiva and Lerman, 1985]. A utility function of this form is defined as

$$U_{ij} = \langle \boldsymbol{\alpha}, \mathbf{Y}_{ij} \rangle \tag{3.32}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_K]$ is a vector of K empirically estimated parameters and $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$ is the dot-product or inner-product of vectors $\mathbf{x}$ and $\mathbf{y}$. Note that linearity in the parameters does not imply that attractiveness is necessarily linear with the attributes $\mathbf{X}_i^P$, $\mathbf{X}_j^F$, and $\mathbf{X}_{ij}^D$. The function $\mathbf{h}$ may be a real transformation, such as a logarithmic or polynomial transformation, of the $\mathbf{X}$ attributes. Fishburn [1970] states that a large number of preference orders can be expressed through a function of this form if the effects of the attributes are independent – the ordering for a given attribute is independent of the levels of the other attributes.

The following example illustrates the use of this choice framework to define an accessibility measure. This example defines a maxitive accessibility measure that is similar to the minimum distance measure and admissible within the generic model.

EXAMPLE 2 *Consider a system with $\mathcal{N} = \{1, \ldots, N\}$ facilities and a given individual i. The feasible choice set for this individual is $\mathcal{A}_i = \{0, \ldots, N\}$. Assume that attraction*

and congestion are unimportant in this situation and are consequently set to a positive value, say 1. Furthermore, suppose that the attractiveness of non-attendance is some negative value, say $U_{i0} = H_i, H_i < 0$, *and that attractiveness decreases linearly with increasing distance. The attractiveness of a given facility might be specified by*

$$U_{ij} = \ln a_{ij} - \ln c_{ij} - \alpha d_{ij} = -\alpha d_{ij}$$

with $\alpha > 0$. *The accessibility, from this choice situation can be expressed as:*

$$\begin{aligned} A_i &= \omega \max(H_i, -\alpha d_{i1}, \ldots, -\alpha d_{iN}) - \omega(-H) \\ &= -\omega H_i - \omega \min(H, \alpha d_{i1}, \ldots, \alpha d_{iN}) \\ &= R - \min(R, d_{i1}, \ldots, d_{iN}) \end{aligned}$$

where $\omega = 1/\alpha$ *and* $R = -H_i/\alpha$, $R > 0$ *are defined for convenience. Thus, this measure is equivalent to the standard minimum distance measure except for the definition of a maximum range,* R, *for a facility. The accessibility of an individual facility is*

$$A_{ij} = \begin{cases} R - D_{ij} & \text{for } D_{ij} \leq R \\ 0 & \text{otherwise.} \end{cases}$$

This measure is consistent with the definition of accessibility in the generic model so that an inaccessible system has an accessibility of zero and accessibility increases with decreasing distance.

Thus, there is a direct link between maxitive accessibility measures and this choice framework. If the attractiveness of a facility is considered to be affected only by its distance from the individual, then the most preferred facility would be the closest and the minimum distance measure is an appropriate accessibility measure. Similar definitions of other maxitive measures, such as the coverage measure, are also possible. With complete knowledge of all the factors affecting the individual's decision-making process, it would be possible to predict exactly which facility the individual chooses and to calculate the accessibility of the system for a particular individual. However, if complete knowledge is unavailable then it is often useful to incorporate a random error term into the measurement of attractiveness. Models of these sorts are termed *random*

utility[4] models and are discussed in the next section.

3.3 Accessibility Measures based on Random Utility Models

The previous section outlined a framework for modelling individual choice behaviour. This framework assumes that an individual chooses the alternative with the highest level of perceived attractiveness. This approach requires complete knowledge of all relevant attributes affecting the choice. However this complete knowledge is unavailable due to observation deficiencies and thus the perceived attractiveness is treated, by the model, as a random variable. Ben-Akiva and Lerman [1985] identify four possible sources of randomness: unobserved attributes of the alternative, unobserved attributes of the individual, measurement errors, and the use of instrumental variables. Therefore, the perceived attractiveness of an alternative to an individual can be considered as the sum of the measured (or observed) attractiveness and the random error term (the unobserved attractiveness). The perceived attractiveness of alternative $j \in \mathcal{A}_i$ to individual i is equal to

$$U_{ij} = V_{ij} + \varepsilon_{ij} \tag{3.33}$$

where V_{ij} is the corresponding measured attractiveness and ε_{ij} is a random error term. For the null alternative, define

$$U_{i0} = V_{i0} + \varepsilon_{i0} = H_i + \varepsilon_{i0} \tag{3.34}$$

where H_i is the attractiveness of non-attendance.

The individual chooses the alternative that has the highest level of attractiveness. However, due to the random terms, the actual perceived attractiveness of each alternative is unknown. Moreover, the total satisfaction or the maximum perceived attractiveness among the alternatives is also a random value. The probability that individual i

[4]The term "random utility" is retained for consistency with the existing literature, for example [Ben-Akiva and Lerman, 1985].

chooses alternative j, equation (3.31), is given by

$$p_{ij} = \Pr(U_{ij} \geq U_{ik}, \forall k \in \mathcal{A}_i) \tag{3.35}$$

$$= \Pr\left[V_{ij} + \varepsilon_{ij} \geq \max_{k \in \mathcal{A}_i}(V_{ik} + \varepsilon_{ik})\right]. \tag{3.36}$$

Furthermore, the satisfaction is

$$U_i^* = \max_{j \in \mathcal{A}_i}\left(V_{ij} + \varepsilon_{ij}\right). \tag{3.37}$$

Any relevant choice model can be derived from equation (3.36). However, in practice, it is difficult to find probability distributions in which these equations have closed-form solutions [Domencich and McFadden, 1975]. One convenient solution results from the assumption that the random error terms have a Gumbel distribution and are independently and identically distributed (IID). The Gumbel distribution is used because it is analytically convenient and approximates a normal distribution.

If a random error term ε is Gumbel-distributed then the cumulative distribution function is $F(\varepsilon) = \exp\left[-e^{\mu(\varepsilon-\eta)}\right], \mu > 0$ and the probability density function is $f(\varepsilon) = \mu e^{-\mu(\varepsilon-mu)} \exp\left[-e^{\mu(\varepsilon-\eta)}\right]$ where η is a location parameter and μ is a positive scale parameter. Furthermore, this distribution has the following properties [Johnson and Kotz, 1970]:

1. The mode of this distribution is η, the mean is $\eta + \gamma/\mu$ where γ is the Euler constant (≈ 0.577), and the variance is $\pi^2/6\mu^2$.
2. If ε is Gumbel-distributed with parameters (η, μ) and a and b are scalar constants, then $a\varepsilon + b$ is Gumbel-distributed with parameters $(a\eta + b, \mu/a)$.
3. If ε_1 and ε_2 are Gumbel-distributed with parameters (η_1, μ) and (η_2, μ) respectively, then $\varepsilon_2 = \varepsilon_1 - \varepsilon_3$ is a logistic distribution with

$$F(\varepsilon_3) = \frac{1}{1 + e^{\mu(\eta_2 - \eta_1 - \eta_3)}}.$$

4. If $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ are n independent Gumbel-distributed random variables with parameters (η_1, μ), $(\eta_2, \mu), \ldots, (\eta_n, \mu)$ respectively, then $\max(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ is Gumbel-distributed with parameters $\left(1/\mu \ln \sum_{j=1}^{n} e^{\mu\eta_j}, \mu\right)$.

For further discussion of the properties of the Gumbel distribution as applied to random utility models, see Ben-Akiva and Lerman [1985, p. 104], Domencich and McFadden [1975, p. 61].

If the random error terms, ε_{ij}, are Gumbel-distributed and IID[5] then the satisfaction is

$$U_i^* = (1/\mu) \ln \sum_{j \in \mathcal{A}_i} \exp \mu V_{ij} + \varepsilon_i^* \tag{3.38}$$

through the use of property 4 of the Gumbel distribution. Furthermore, the expected value of the satisfaction is simply

$$V_i^* = E(U_i^*) = (1/\mu) \ln \sum_{j \in \mathcal{A}_i} \exp \mu V_{ij} + \gamma/\mu \tag{3.39}$$

where γ is the Euler constant. Furthermore, it can be shown [Domencich and McFadden, 1975; Daganzo, 1979] that the choice probability, equation (3.31), is:

$$p_{ij} = \frac{\exp \mu V_{ij}}{\sum_{k \in \mathcal{A}_i} \exp \mu V_{ik}}. \tag{3.40}$$

Equation (3.40) defines the multinomial logit model [Domencich and McFadden, 1975], which has been used extensively to estimate empirical choice probabilities in a variety of applications. Moreover, it is possible to simplify these equations by scaling the vector of parameters, α, defining U_{ij} so that the scale parameter, μ, is equal to one. Finally, recall that in equation (3.28), accessibility is defined as some monotonic function of the difference between the satisfaction of the choice set and the satisfaction of a system consisting of only the null alternative. As both of these satisfaction values are shifted by a constant, γ/μ, this constant does not affect accessibility and, therefore, can be omitted.

The form of the accessibility measure depends upon the monotonic function that transforms satisfaction into accessibility. As discussed previously, most existing accessibility measures are defined using either a linear or an exponential transformation function. If an exponential transformation function is used, then the accessibility is

$$A_i = \sum_{j \in \mathcal{A}_i} \exp V_{ij} - \exp H_i = \sum_{j \in \mathcal{N}_D} V_{ij}. \tag{3.41}$$

[5]The location parameter is also assumed to be zero. However, this parameter can be absorbed into the measured attractiveness.

It is interesting to note that this equation is independent of the value of the attractiveness of non-attendance, H_i. On the other hand, if a linear transformation is applied to (3.39), then accessibility is defined by

$$A_i = \ln\left(\exp H_i + \sum_{j\in\mathcal{N}_\mathcal{D}} \exp V_{ij}\right) - H_i \quad . \tag{3.42}$$

For a linear transformation, the attractiveness of non-attendance does contribute to the accessibility measure. For example, if H_i is defined to be zero, then equation (3.42) is

$$A_i = \ln\left(1 + \sum_{j\in\mathcal{N}_\mathcal{D}} \exp V_{ij}\right) \tag{3.43}$$

which is a transform-separable accessibility measure within the generic model.

The following example illustrates the derivation of a gravity-model accessibility measure using the random utility framework. In this example, the Joseph and Bantock [1982] accessibility measure (discussed in Section 2.1.2) with an exponential distance decay function is derived.

EXAMPLE 3 *Consider a system with $\mathcal{N} = \{1, \ldots, N\}$ facilities and a given individual i. The feasible choice set for this individual is $\mathcal{A}_i = \{0, \ldots, N\}$. Assume that the attractiveness of non-attendance, H_i, is 0 and that the attractiveness of a facility j is equal to its size, $a_{ij} = S_j$. Furthermore, the congestion of a facility is defined as follows*

$$c_{ij} = C_j = \sum_{i\in\mathcal{N}_\mathcal{I}} \exp\left(-\beta D_{ij}\right)$$

so that facilities located near large populations have a high congestion. Finally, let the attractiveness decrease linearly with increasing distance with a slope of β. Thus, the measured attractiveness of a given facility can be specified by

$$V_{ij} = \ln a_{ij} - \ln c_{ij} - d_{ij} = \ln S_j - \ln C_j - \beta D_{ij}.$$

Using an exponential transformation, the accessibility is

$$A_i = \sum_j \left(S_j / C_j\right) \exp\left(-\beta D_{ij}\right).$$

This equation is identical to the Joseph and Bantock accessibility measure defined in (2.5) *using an exponential distance decay, except that it is expressed in an unaggre-*

gated form. An individual facility's attractiveness is

$$A_{ij} = S_j / C_j \exp\left(-\beta D_{ij}\right).$$

Using a linear transformation, accessibility can also be expressed as

$$A_i = \ln\left[1 + \sum_j S_j / C_j \exp\left(-\beta D_{ij}\right)\right].$$

One important property of the multinomial logit formulation is *independence from irrelevant alternatives* (IIA) so that "the probability of choice of two alternatives depends only on their measured attractiveness" [Daganzo, 1979, p. 10]. This is relatively easy to verify as

$$p_{ij} / p_{ik} = \exp V_{ij} / \exp V_{ik}. \tag{3.44}$$

Since the random error terms are independent, there is no correlation between the errors and, therefore, the perceived attractiveness values. Thus, one alternative does not influence the perceived attractiveness of any other alternative. Thus, the accessibility measure defined by this process is either strictly separable or transform-separable. For example, equation (3.41) is strictly separable while equation (3.42) is transform-separable. The next section discusses a model formulation where the perceived attractiveness values can be correlated.

3.4 Random Utility Models with Correlated Alternatives

As discussed previously, the assumption of IID Gumbel-distributed random error terms leads to a multinomial logit formulation for the choice model. An important assumption of this model is that the random error terms are mutually independent. If these random variables are correlated due to a shared unobserved component of the attractiveness of alternatives, then the assumption of independence would be inappropriate and the previous formulation would provide incorrect results.

Often individuals may perceive groups or classes of alternatives as being similar. In the context of health care accessibility, a reasonable classification of facilities would be by their type or level in the health care hierarchy – hospitals, clinics, and health posts. Thus, due to these shared unobserved components of attractiveness, the random error

terms are no longer independent and the perceived attractiveness of alternatives within a class would be correlated. This model is termed the nested logit formulation.

Suppose that the alternatives are grouped into K mutually exclusive and collectively exhaustive classes, with C_k being the set of alternatives in class k. Now consider the perceived attractiveness, U_j, of a facility in class k[6]. Suppose that the facilities in a cluster share some observed and unobserved attributes. Consequently, the perceived attractiveness of an alternative j in cluster k can be written as

$$U_j = U_k^C + U_j^A = V_k^C + V_j^A + \varepsilon_k^C + \varepsilon_j^A \tag{3.45}$$

where U_k^C, U_j^A are the perceived attractiveness terms, V_k^C, V_j^A are the measured attractiveness, and $\varepsilon_k^C, \varepsilon_j^A$ are the random error terms for class k and alternative j respectively.

If the random error terms of alternatives within a cluster, ε_j^A, are IID Gumbel-distributed with a scale parameter of μ_k, then the satisfaction is

$$\begin{aligned} U^* &= \max_{j \in \mathcal{A}} U_j \\ &= \max_k \left[V_k^C + \varepsilon_k^C + \max_{j \in S_k} \left(V_j^A + \varepsilon_j^A \right) \right] \\ &= \max_k \left(V_k^C + \varepsilon_k^C + U_k' \right) \end{aligned} \tag{3.46}$$

where U_k' is the resulting satisfaction of the facility-dependent attractiveness for all facilities in cluster k. However, since the attractiveness values of all the facilities in cluster k are IID Gumbel-distributed with scale parameter, μ_k, then

$$U_k' = (1/\mu_k) \ln \textstyle\sum_{j \in S_k} \exp \mu_k V_j^A + \varepsilon_k' = V_k' + \varepsilon_k' \quad . \tag{3.47}$$

Thus

$$U^* = \max_k \left(V_k^C + V_k' + \varepsilon_k^C + \varepsilon_k' \right) . \tag{3.48}$$

If ε_k^C is distributed so that $\varepsilon_k = \varepsilon_k^C + \varepsilon_k'$ is Gumbel-distributed with a scale parameter of

[6]The following discussion omits the subscript i denoting the individual for the sake of notational clarity. It should be clear where this subscript is implicit.

μ_C, then[7]

$$V^* = E(U^*) = (1/\mu_C) \ln \sum_k \exp \mu_C \left(V_k^C + V_k' \right)$$
$$= (1/\mu_C) \ln \sum_k \left[\exp \mu_C V_k^C \left(\sum_{j \in C_k} \exp \mu_k V_j^A \right)^{\mu_C/\mu_k} \right]. \tag{3.49}$$

With appropriate scaling chosen so that $\mu_C = 1$ and defining $\theta_k = \mu_k/\mu_C$ equation (3.49) becomes

$$V^* = \ln \sum_k \left[\exp V_k^C \left(\sum_{j \in C_k} \exp \theta_k V_j^A \right)^{1/\theta_k} \right] \tag{3.50}$$

where $\theta_k \geq 1$.

Furthermore, θ_k also has a very natural interpretation in terms of the correlation of the perceived attractiveness between any two alternatives in the same cluster [Ben-Akiva and Lerman, 1985], namely,

$$\text{Corr}\left(U_j, U_j'\right) = 1 - \theta_k^{-2} \tag{3.51}$$

for $j, j' \in C_k$. Values of θ_k less than one do not have a rational interpretation within random utility framework [Fotheringham and O'Kelly, 1989]. If $\theta_k = 1$, then there is no correlation between alternatives in that cluster and if every $\theta_k = 1$ then the model is equivalent to the multinomial logit model.

EXAMPLE 4 *Consider a system with three types of facilities: hospitals, clinics, and health posts. Let C_k be an index set of each type facility. Suppose that facilities of the same type are perceived by the individual as similar and the correlation between facilities of type k is $1 - \theta_k^{-2}$ and the measured attractiveness is $V_{ik}^C = \lambda_k$. Define the measured attractiveness of a facility j in C_k as before with*

$$V_{ij}^A = \ln a_{ij} - \ln c_{ij} - d_{ij} = \ln S_j - \ln C_j - \beta_k D_{ij}$$

where λ_k is the perceived attractiveness of a facility of type k and the distance decay parameter, β_k, varies by facility type. Furthermore, the null alternative is in its own clus-

[7]The constant term, γ/μ, has been omitted in these expressions.

ter with an attractiveness of H_i. Consistent with the previous example, the congestion is defined by

$$c_{ij} = C_j = \sum_{i \in \mathcal{N}_d} \exp -\beta_k D_{ij}.$$

Using an exponential transformation, the accessibility is

$$A_i = \sum_{k=1}^{3} L_k \left[\sum_{j \in S_k} \left(S_j / C_j \right)^{\theta_k} \exp \left(-\beta_k \theta_k D_{ij} \right) \right]^{1/\theta_k}$$

where $L_k = \exp \lambda_k$ is the relative attractiveness of a facility of type k. Similarly, the accessibility can also be expressed as

$$A_i = \ln \left\{ \exp H_i + \sum_{k=1}^{3} L_k \left[\sum_{j \in S_k} \left(S_j / C_j \right)^{\theta_k} \exp \left(-\beta_k \theta_k D_{ij} \right) \right]^{1/\theta_k} \right\} - H_i.$$

The previous example defines an aggregable accessibility measure but not a separable one. Since the random error terms for each facility are independent from the null alternative, the accessibility (using an exponential transformation) of an individual facility is

$$A_{ij} = L_k S_j / C_j \exp \left(-\beta_k D_{ij} \right). \tag{3.52}$$

Thus, the overall accessibility can be written as

$$A_i = \sum_k \left(\sum_{j \in S_k} A_{ij}^{\theta_k} \right)^{1/\theta_k} \tag{3.53}$$

and is commutative, monotonic, and has a zero identity element, and would be zero if there were no facilities. Hence, this defines an aggregable accessibility measure. Furthermore, if $\theta_k = \theta$ is the same for every cluster, then it is possible to define the aggregation operator as

$$X \oplus Y = \left(X^{\theta} + Y^{\theta} \right)^{1/\theta} \tag{3.54}$$

so that the measure would be separable. However, if θ_k are not all equal, then the measure does not have a separable representation.

3.5 Summary

This chapter introduced a generic model for potential accessibility measures. The measures defined by this model are aggregable, *i.e.*, it is possible to define the accessibility of an individual to a facility in the system and combine these facility accessibility values using an aggregation operator to form a measure of overall system accessibility. For an individual, the accessibility to a facility is proposed to be a function of the attraction of the facility, the distance between the individual and the facility, and the congestion of the facility. Furthermore, attraction has a positive linkage with accessibility while congestion of a facility affects accessibility negatively as does the distance between an individual and a facility if the distances is greater than some minimum range. If the aggregation operator is associative, then the accessibility measure is considered strictly separable so that there is no interaction between facilities. Two well known cases of strictly separable measures are maxitive accessibility measures and additive accessibility measures. An accessibility measure is transform-separable if there is a monotonic function which transforms the measure into a strictly separable measure.

The generic model does not provide any behavioural justification for the definition of an accessibility measure. Therefore, this chapter also presented a behavioural framework for the generic model, based on individual choice theory. In this framework, the individual chooses the most attractive alternative out of the facilities in the system augmented by the "null" alternative – the decision to choose no facility. The accessibility of the system is a monotonic function of the "satisfaction", or the expected attractiveness of the chosen alternative, of the system scaled so that the accessibility of a system consisting of only the null alternative is zero. This results in a maxitive accessibility measure if attractiveness is a deterministic quantity and, hence, known exactly.

Often, due to observational deficiencies and imprecise data, attractiveness is modelled as a stochastic variable and partitioned into a measured attractiveness component and a random error term. Additive accessibility measures, corresponding to the familiar gravity modelling approach, result when the random error terms are identically and independently Gumbel-distributed. If the assumption of independence were violated due to shared unobserved factors between facilities, then a modified accessibility measure that is no longer separable resulted.

The models discussed in this chapter focus exclusively on examining the accessibility of a single individual to a system of facilities. However, a goal of accessibility mea-

sures is to evaluate the accessibility of the total population of potential users. Thus, the accessibility levels of individuals must be combined to form aggregate-level measures of accessibility. The process of aggregation is discussed in the next chapter.

Chapter 4

The Effects of Aggregation on Accessibility Measures

The effect of aggregation[1] is a problem that permeates all geographic measures of health care accessibility. The generic model discussed in the previous chapter is defined in terms of the accessibility of individuals to health care facilities. This accessibility is influenced by relevant individual characteristics, the characteristics of the facilities, and the spatial separation between individuals and facilities. However, often we are interested in examining the accessibility of different population *groups* and different *geographic regions* within a study area. Conceptually, the individual-level accessibility measures are combined into aggregate measures for groups of individuals.

This chapter examines the process of measuring the accessibility of groups of individuals. The first section outlines the process of aggregation and considers both the aggregation of individual characteristics and the spatial aggregation of individuals. Next, several worst-case bounds are derived for spatial aggregation error. These bounds are derived for both a minimum-distance accessibility measure and for a gravity-type measure with exponential distance deterrence. The level of spatial aggregation error present in population zones can be reduced if the population is disaggregated to a raster or grid data set of a given spatial resolution. The final section in this chapter outlines a method developed by Bracken and Martin [1989] for disaggregation from population points to

[1]Not to be confused with an aggregable accessibility measure – a measure in which the accessibility levels of an individual consumer to a single facility are combined to form an overall accessibility level *of an individual* to a system of facilities.

a grid. Two extensions are proposed for this method to incorporate land use classifications and for disaggregation from areas to grid cells.

4.1 The Process of Aggregation

Consider a sub-group of P individuals from the population of potential users of health facilities and define an index set $\mathcal{P} \subseteq \mathcal{N}_{\mathcal{U}}$ for this group. The aggregated accessibility of this sub-group is the average accessibility of all the individuals to the facilities. Thus, the aggregate accessibility is

$$A = (1/P) \sum_{i \in \mathcal{P}} A\left(\mathbf{X}_i^P, \mathcal{F}, \mathbf{D}_i\right) \tag{4.1}$$

where $\mathbf{X}_i^P$ is a vector of attributes of individual i, $\mathcal{F}$ is the set of facility attributes, and $\mathbf{D}_i$ is a vector of distances between this individual and the facilities. Conceptually, it is a straightforward process to calculate an aggregate accessibility measure: simply evaluate the accessibility for each individual and then average these accessibility values.

However, the use of equation (4.1) requires complete knowledge of the vector of relevant characteristics and the exact location for each individual. While it may be possible to obtain this information for highly focussed studies of revealed accessibility or utilization, it is generally impossible, especially in a developing country with limited data resources, to obtain a complete database of this information. This is particularly true for studies of potential accessibility where the number of users can be very large. In addition, data on individuals are often only available at an aggregated level due to reasons of confidentiality or the cost of data storage [Hodgson and Neuman, 1993]. Even if these data were available, it would be computationally infeasible to calculate the aggregate accessibility measure in this manner. Therefore, the goal of aggregation "is to develop methods for reducing the required data and computation" [Ben-Akiva and Lerman, 1985, p. 133] needed to evaluate the accessibility of the system to a large group of individuals.

One possible aggregation strategy is to construct an "average individual," estimated from the characteristics and spatial distribution of the population [Ben-Akiva and Lerman, 1985]. Noting that the facility characteristics remain constant for all individuals,

equation (4.1) becomes

$$\tilde{A} = A\left(\tilde{\mathbf{X}}^P, \mathcal{F}, \tilde{\mathbf{D}}\right) \tag{4.2}$$

where $\tilde{\mathbf{X}}^P$ is a vector of the average individual characteristics and $\tilde{\mathbf{D}}$ is a vector of the average distance of an individual to each facility. The overall aggregation error would be

$$E^A = A - \tilde{A} = (1/P)\sum_{i\in\mathcal{P}} A\left(\mathbf{X}_i^P, \mathcal{F}, \mathbf{D}_i\right) - A\left(\tilde{\mathbf{X}}^P, \mathcal{F}, \tilde{\mathbf{D}}\right). \tag{4.3}$$

However, this method has certain properties that must be considered in its application. As Ben-Akiva and Lerman [1985] note, the aggregation error increases with increasing variance in the distribution of the characteristics so that for heterogeneous populations aggregation error can become substantial and approach 100% error in populations with very large variances. Moreover, the use of this measure can mask important variations in accessibility among individuals.

One way of reducing the effect of these properties is to use the technique of classification [Ben-Akiva and Lerman, 1985]. In general, the error is greatest when the distribution of the variables being aggregated has a high variance. The classification method reduces this variance by grouping the population into a set of relatively homogeneous sub-groups and applying the average individual technique to each sub-group. This method can be specified as follows.

1. Partition the population of potential users into K mutually exclusive and collectively exhaustive sub-groups. The goal of this partitioning is to select population sub-groups whose characteristics are similar so that the variance in their characteristics and distances, and hence, their accessibility is relatively small. In other words, the population is partitioned so that between group variance is maximized and within group variance is minimized.

2. Calculate (or estimate) the number of users in each of the sub-groups, P_k.

3. For each group, select a representative value for their characteristics, $\tilde{\mathbf{X}}_k^P$, and the distances, $\tilde{\mathbf{D}}_k$.

4. Calculate the aggregate accessibility as the weighted average of the accessibility

of each sub-group, *i.e.*,

$$\tilde{A} = \sum_{k=1}^{K} \frac{P_k}{P} A\left(\tilde{\mathbf{X}}_k^P, \mathcal{F}, \tilde{\mathbf{D}}_k\right). \tag{4.4}$$

The classification method has a further advantage in that it allows for the comparison of accessibility levels between groups. This information can be extremely valuable in planning situations where it is important to examine differential accessibility among different regions or social groups.

From equation (4.4), there are two main sources of aggregation error: errors caused by the aggregation of the population's demographic characteristics, and errors in distance estimates caused by the spatial aggregation of the individuals. This suggests that the individuals should be partitioned by their characteristics, so that the variance in $\tilde{\mathbf{X}}_k^P$ is minimized, and spatially to reduce the error in the distance estimates. The error induced by the aggregation of individual characteristics is less important for potential geographic accessibility measures as most of these measures utilize very few demographic characteristics. For example, in the review of potential geographic accessibility measures in Chapter 2, only the measure proposed by Knox [1978; 1979] uses population characteristics, and it uses only car ownership. However, Ben-Akiva and Lerman [1985] note that the classification method, with demographic characteristics, works well even when a small number of classes is used. Furthermore, they note that "empirical evidence suggests that the errors due to aggregation across individuals can be made relatively small without a great deal of difficulty" (p. 153). In contrast, the issue of the spatial aggregation of individuals is more complicated. The next section addresses this issue.

4.2 Spatial Aggregation of Individuals

An important source of distance estimation error is due to spatial aggregation error. Recall from the previous section that the goal of the aggregation process is to partition the population so as to reduce the within group variation of their attributes. Obviously, when aggregating individuals spatially, locations that are near each other have similar distances to the facilities. When individuals are spatially aggregated the population groups are defined as geographic sub-areas and the individuals in these areas are as-

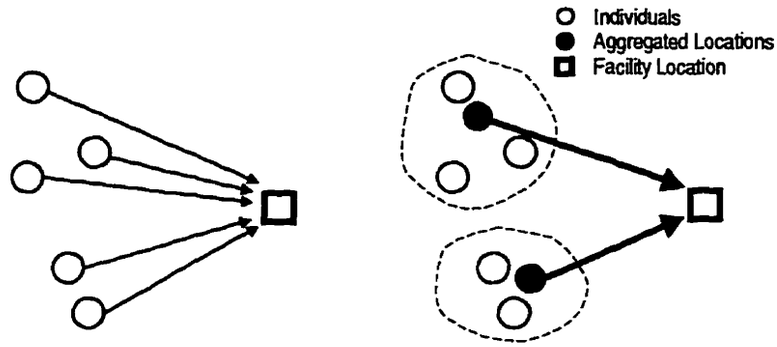


Figure 4.1: The spatial aggregation process.

sumed to be at single locations, termed the aggregate locations, so that $\mathbf{D}_i = \widetilde{\mathbf{D}}_k$ for all individuals i in group k. Typically this aggregate location is assumed to be at the geographic centre of often irregularly-shaped administrative areas or districts. The spatial aggregation process is illustrated in Figure 4.1.

Distance errors are caused by the fact that, in general, the distance from an individual to a facility is not equal to the distance between the corresponding aggregate location and the same facility. Assuming that every individual in the group has the same characteristics, then the average error in accessibility for group k is defined by

$$E_k^A = (1/P_k) \sum_{i \in \mathcal{P}_k} A\left(\mathbf{X}_k^P, \mathcal{F}, \mathbf{D}_i\right) - A\left(\mathbf{X}_k^P, \mathcal{F}, \widetilde{\mathbf{D}}_k\right). \tag{4.5}$$

The specific distance errors from spatial aggregation depend upon the form of the accessibility measure. Moreover, this quantity is also dependent on both the spatial distribution of the individuals in the population group and the relative location of the facilities. Thus, the level of distance error caused by the spatial aggregation of the individuals cannot, in general, be described with a closed form equation and is dependent upon the particular situation. However, it is possible to develop worst-case error bounds for some common accessibility measures. The next two subsections examine spatial aggregation induced error for the minimum distance measure and for the gravity-model measure.

4.2.1 Spatial Aggregation Error for Minimum Distance Accessibility Measures

The spatial aggregation process can cause substantial errors in the distance estimation, and hence accessibility, from the population to the system of facilities [Hillsman and Rhoda, 1978]. While this problem has been extensively studied in location-allocation modelling [Goodchild, 1979; Casillas, 1987; Fotheringham *et al.*, 1995], the effects of spatial aggregation of a population have largely been ignored in operational studies of potential health care accessibility. There is, however, a strong relationship between location-allocation modelling, particularly the p-median model, and the minimum distance accessibility measure[2]. In fact, the objective of the p-median model is to locate p facilities so as to minimize the average distance between the demand (population) and the supply nodes (facilities). This is similar to finding the most accessible configuration of facilities using the minimum distance measure. Thus, the studies of aggregation error for location-allocation models are equally applicable for the minimum distance accessibility measures.

As noted previously, the spatial aggregation of individuals induces errors in the estimates of aggregate-level accessibility. This section examines this process using the minimum distance accessibility measure. For this measure, Hillsman and Rhoda [1978] categorized aggregation error into three different sources with two different causes. Source A errors arise because the distance from an individual to a facility is not equal to the distance between the corresponding aggregate location and the same facility. Source B errors are a special case of source A errors and occur when a potential facility site is at the same position as an aggregate. Source B errors cause the aggregate accessibility to be higher than the actual accessibility while source A errors can cause either an underestimation or an overestimation of accessibility. Source C errors occur because all individuals in a given population zone are allocated to a single facility even though some of the individuals may be closer to a different facility. Source C errors cause the aggregate accessibility measure to underestimate the actual accessibility of the individuals. These three sources of spatial aggregation error are illustrated in Figure 4.2 and are associated with error due to incorrect distance estimation and in the calculation of the nearest facility.

[2]As per Example 2 in the previous chapter, the minimum distance measure can be considered as a maxitive accessibility measure within the generic model of potential accessibility. This is discussed further

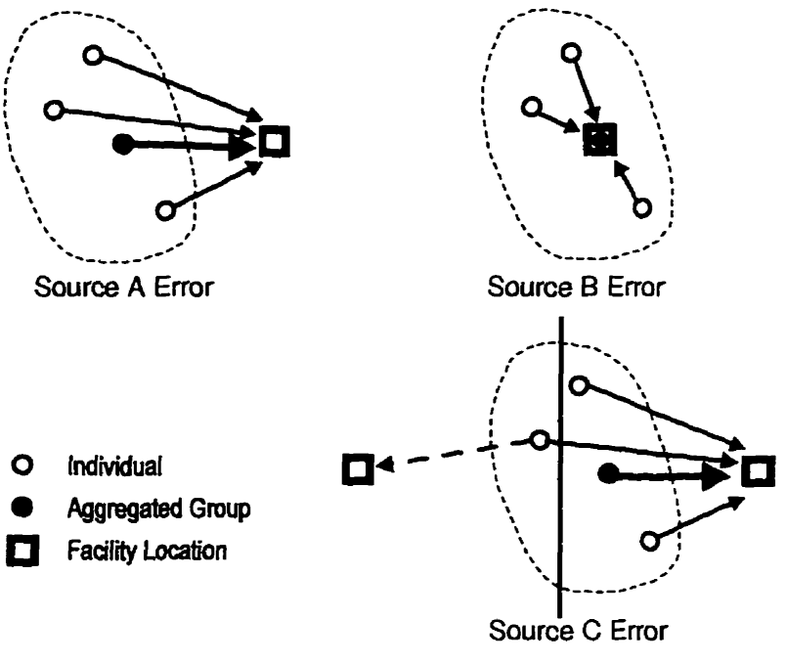


Figure 4.2: Sources of spatial aggregation error.

The effects of aggregation on the measurement of the minimum distance to a set of facilities is recognized by many authors. Goodchild [1979, p. 249] identifies the importance of the spatial aggregation process and further notes that the "effects of aggregation are unique to particular [situations], and therefore ... no general rules of aggregation can be found." In a recent study, Fotheringham *et al.* [1995] found that solutions to the *p*-median problem were highly sensitive to both the level of aggregation and the definitions of the geographic sub-areas; the results "obtained from such an analysis [pertain] to a specific set of zonal demand data and not necessarily to the true underlying demand structure" (p. 74). Francis and Lowe [1992, p. 232] also recognize that "too much aggregation can destroy the accuracy" of measuring the minimum distance. Furthermore, they show that determining an aggregation scheme which causes the least error (in the worst case sense) in the measurement of the minimum distance is an *NP*-hard problem. Therefore, they validate Goodchild's [1979, p. 249] claim of there being "no general rules of aggregation."

However, Francis and Lowe [1992] demonstrate that it is possible to derive a worst-

in Chapter 6.

case estimate for spatial aggregation error in the minimum distance accessibility measure. Consider the accessibility measure discussed in Example 2 of Section 3.2 of the previous chapter. In order to simplify the analysis, assume that $R > \min_j D_{ij}$ so that all individuals are always within range of a facility and hence "covered" by the system. The accessibility for a given individual i is defined by

$$A_i = R - \min_j D_{ij}. \tag{4.6}$$

Substituting this definition into equation (4.5), the spatial aggregation error is

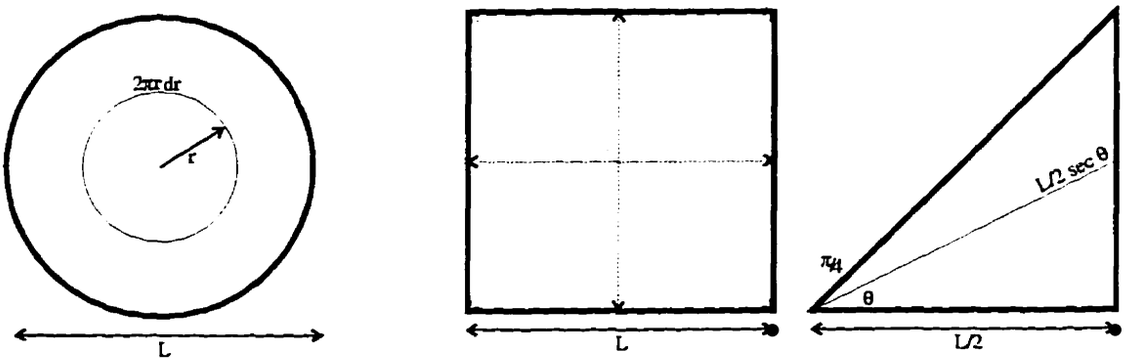
$$\begin{aligned} E_k^A &= (1/P_k) \textstyle\sum_{i \in \mathcal{P}_k} (R - \min_j D_{ij}) - (R - \min_j D_{kj}) \\ &= \min_j D_{kj} - (1/P_k) \textstyle\sum_{i \in \mathcal{P}_k} \min_j D_{ij}. \end{aligned} \tag{4.7}$$

It is possible to derive a worst-case bound on the spatial aggregation error assuming that the distance has the customary properties of a distance measure, namely $D_{ij} = D_{ji}$ (symmetry), $D_{ij} \geq 0$ with $D_{ij} = 0$ implying that $i = j$ (nonnegativity), and $D_{kj} \leq D_{ki} + D_{ij}$ (the triangle inequality). Since, through the use of the triangle inequality, $D_{ij} \leq D_{ik} + D_{kj}$, an upper bound for the maximum aggregation error is

$$\begin{aligned} E_k^A &= \min_j D_{kj} - (1/P_k) \textstyle\sum_{i \in \mathcal{P}_k} \min_j D_{ij} \\ &\leq \min_j D_{kj} - (1/P_k) \textstyle\sum_{i \in \mathcal{P}_k} (D_{ik} + \min_j D_{kj}) \\ &= (1/P_k) \textstyle\sum_{i \in \mathcal{P}_k} D_{ik}. \end{aligned} \tag{4.8}$$

An upper bound on the error in measuring accessibility caused by spatially aggregated individuals is simply the average distance of the individuals to the aggregate location. The lower bound on the aggregation error can be found using $D_{kj} \leq D_{ik} + D_{ij}$ and is the negative of (4.8). Note that this bound exactly describes the aggregation error only when the triangle inequality is an equality. For other cases, depending on the spatial pattern of the individuals and the facilities, this bound can overestimate the level of aggregation error. The worst-case bound for the minimum distance accessibility measure for a given population sub-group k is

$$\left| E_k^A \right| \leq (1/P_k) \textstyle\sum_{i \in \mathcal{P}_k} D_{ik}. \tag{4.9}$$



(a) Circular Sub-Area (b) Square Sub-Area (c) Integration Detail

Figure 4.3: Calculating E_k^A for *(a)* circular and *(b)* square sub-areas.

Thus, the problem of choosing a set of aggregate locations to minimize this bound is a p-median problem and, therefore, NP-hard.

For certain situations it is possible to derive a closed form expression for this error bound. Three specific cases for which this bound can be computed are: a circular sub-area of diameter L using straight-line distances (ℓ_2 norm), a square sub-area with sides of length L using straight-line distances, and the square sub-area using rectilinear distances (ℓ_1 norm). Each of these cases assumes a uniform population density, ρ_k, defined by the total population divided by the area so that $\rho_k = P_k/a$, where a is the area. An alternate interpretation is that the population is randomly and uniformly distributed over the sub-area and that we are calculating the expected value of the distance to the aggregate location. For each of these cases, the aggregation error bound is equal to

$$\left|E_k^A\right| \leq \frac{1}{P_k}\int_\Gamma \rho_k r da = \frac{1}{a}\int_\Gamma r da \tag{4.10}$$

where Γ is the respective sub-area.

Circular sub-area – Straight-line distances Consider the circular sub-area shown in Figure 4.3a. The bound on the aggregation error in this situation has a particularly simple form. Integrating the sub-area using circular strips, equation (4.10) can be expressed as follows

$$\left|E_k^A\right| \leq \frac{4}{\pi L^2}\int_0^{L/2} r 2\pi r dr = \frac{1}{3}L. \tag{4.11}$$

Thus, the expected value of the aggregation error for the case of circular sub-areas is one-third the diameter of the circle.

Square sub-area – Straight-line distances The square sub-area, shown in Figure 4.3b, can be partitioned into 8 equal triangles. For each of these triangles, Figure 4.3c, the integral in equation (4.10) can be expressed in polar coordinates as

$$\int_0^{\pi/4} \int_0^{L/2\sec\theta} r^2 dr d\theta$$

so that the error bound can be expressed as

$$\begin{aligned} \left|E_k^A\right| &\leq \frac{8}{L^2} \int_0^{\pi/4} \int_0^{L/2\sec\theta} r^2 dr d\theta \\ &= \frac{1}{6}\left[\sqrt{2} + \ln\left(\sqrt{2}+1\right)\right] \approx 0.3826L. \end{aligned} \tag{4.12}$$

Square sub-area – Rectilinear distances The distance between $x = (x_1, x_2)$ and $y = (y_1, y_2)$ using the ℓ_1 norm (rectilinear distance) is defined as

$$D_{xy} = |y_1 - x_1| + |y_2 - x_2|. \tag{4.13}$$

If the square is sub-divided into four equal quadrants, the aggregation error bound using rectilinear distances can be expressed as

$$\begin{aligned} \left|E_k^A\right| &\leq \frac{4}{L^2} \int_0^{L/2} \int_0^{L/2} (x+y) dx dy \\ &= L/2. \end{aligned} \tag{4.14}$$

Thus, for each of these cases, the spatial aggregation error bound increases linearly with the length of the spatial units. Furthermore, these expressions provide an estimate of the potential error caused by using aggregate-level accessibility measures.

It is important to note that error bounds for the aggregation error are independent of the actual level of accessibility. For a particular sub-area, it is possible to calculate the maximum possible aggregation error. For example, consider a situation using straight-line distances with square sub-areas. Suppose that accessibility is being measured with equation (4.6) with the maximum distance, R, being 20 kilometres.

	Error	Minimum Distance					
Length (L)	Bound (E^A)	0	0.5	1	2	5	10
0.25	0.096	0.48%	0.49%	0.50%	0.53%	0.64%	0.96%
0.5	0.191	0.96%	0.98%	1.01%	1.06%	1.28%	1.91%
1	0.383	1.91%	1.96%	2.01%	2.13%	2.55%	3.83%
2	0.765	3.83%	3.92%	4.03%	4.25%	5.10%	7.65%
4	1.530	7.65%	7.85%	8.05%	8.50%	10.2%	15.3%

Table 4.1: Table of relative aggregation error bounds for minimum distance accessibility measures with a maximum distance, R, of 20 kilometres.

Table 4.1 summarizes the error bound and the relative aggregation error for both different-sized sub-areas and for five different minimum distances to the nearest facility. For this particular example, the level of relative aggregation error remains fairly low (less than 16%). This is due to the addition of the constant R to the accessibility measure which tends to reduce the effect of the aggregation error for high levels of accessibility. Correspondingly, the effect of aggregation error is magnified at low levels of accessibility. At a minimum distance of 19 (hence an accessibility of 1), if $L = 4$ the possible aggregation error is greater than the accessibility. Although it is important to consider the effects of spatial aggregation error on minimum-distance accessibility measures, gravity-type accessibility measures are much more susceptible to these effects. This is discussed in the next section.

4.2.2 Spatial Aggregation Error of Gravity Model Accessibility Measures

Using a similar technique to that presented in the previous section, it is possible to derive worst-case error bounds for gravity model accessibility errors that use an exponential distance deterrence function. Recall the standard Hansen accessibility measure discussed in Chapter 2. The accessibility for a given individual i is defined by

$$A_i = \sum_j S_j \exp\left(-\beta D_{ij}\right). \tag{4.15}$$

Substituting this accessibility into equation (4.5), the spatial aggregation error with this accessibility measure is

$$E_k^A = (1/P_k) \sum_{i \in P_k} \sum_j S_j \exp\left(-\beta D_{ij}\right) - \sum_j S_j \exp\left(-\beta D_{kj}\right). \tag{4.16}$$

Again, assume that D_{ij} has the customary properties of a distance measure. Using the triangle inequality and symmetry, we know that

$$D_{ij} \geq D_{kj} - D_{ik} \tag{4.17}$$

and therefore,

$$\exp\left(-\beta D_{ij}\right) \leq \exp\left[-\beta\left(D_{kj} - D_{ik}\right)\right] \tag{4.18}$$

$$= \exp\left(\beta D_{ik}\right)\exp\left(-\beta D_{kj}\right). \tag{4.19}$$

Note that the sign of the inequality is reversed since $\exp(-\beta D_{ij})$ is a monotonically decreasing function of increasing distance[3]. Therefore, an upper bound for the aggregation error is

$$\begin{aligned} E_k^A &\leq \left(1/P_k\right)\sum_{i\in P_k}\sum_j S_j \exp\left(\beta D_{ik}\right)\exp\left(-\beta D_{kj}\right) - \sum_j S_j \exp\left(-\beta D_{kj}\right) \\ &= \left[\left(1/P_k\right)\sum_{i\in P_k}\exp\left(\beta D_{ik}\right) - 1\right]\sum_j S_j \exp\left(-\beta D_{kj}\right) \\ &= \left(F_k^A - 1\right) A_k. \end{aligned} \tag{4.20}$$

Also note that the maximum aggregation error is within a certain fraction of the level of accessibility. If F_k^A is defined to be the aggregation error fraction, then this error fraction is independent of the characteristics of the facilities and is only a function of the spatial distribution of the individuals with respect to their aggregation points. Using a similar procedure to develop a lower bound for the aggregation error, the aggregation error fraction is constrained to lie within

$$\left(1/P_k\right)\sum_{i\in P_k}\exp\left(-\beta D_{ik}\right) \leq F_k^A \leq \left(1/P_k\right)\sum_{i\in P_k}\exp\left(\beta D_{ik}\right). \tag{4.21}$$

It is possible to derive expressions for the error fraction for the same three specific cases as discussed previously: a circular sub-area of diameter L using straight-line distances (ℓ_2 norm), a square sub-area with sides of length L using straight-line distances, and the square sub-area using rectilinear distances (ℓ_1 norm).

If a population is distributed continuously and uniformly on a circular sub-area of diameter L using straight-line distances, then the population density of this sub-area is

[3]That is, if $a \geq b$, then $exp(-a) \leq \exp(-b)$.

the total population divided by the area. In this case, the bounds for the aggregation error fraction are

$$\frac{4}{\pi L^2}\int_0^{L/2}\exp(-\beta r)2\pi r dr \leq F_k^A \leq \frac{4}{\pi L^2}\int_0^{L/2}\exp(\beta r)2\pi r dr. \tag{4.22}$$

Evaluating these integrals leads to the following bounds

$$4\frac{2-\exp(-\beta L/2)(\beta L+2)}{\beta^2 L^2} \leq F_k^A \leq 4\frac{\exp(-\beta L/2)(\beta L+2)-2}{\beta^2 L^2}. \tag{4.23}$$

For the case of square sub-areas using straight-line distances, a similar strategy to that used for calculating the minimum distance error yields error bounds of

$$\frac{8}{L^2}\int_0^{\pi/4}\int_0^{L/2\sec\theta}\exp(-\beta r)r dr d\theta \leq F_k^A \leq \frac{8}{L^2}\int_0^{\pi/4}\int_0^{L/2\sec\theta}\exp(\beta r)r dr d\theta. \tag{4.24}$$

Unfortunately these integrals do not have closed-form solutions. However, it is possible to find a closed-form solution using the assumption that distance is measured using the ℓ_1 norm (Manhattan distance). If the square is sub-divided into four equal quadrants, then the aggregation error bound using Manhattan distances can be expressed as

$$\frac{4}{L^2}\int_0^{L/2}\int_0^{L/2}\exp\left[-\beta(x+y)\right]dxdy \leq F_k^A \leq \frac{4}{L^2}\int_0^{L/2}\int_0^{L/2}\exp\left[\beta(x+y)\right]dxdy. \tag{4.25}$$

Evaluating these integrals yields

$$4\frac{\exp(-\beta L)-2\exp(-\beta L/2)+1}{\beta^2 L^2} \leq F_k^A \leq 4\frac{\exp(\beta L)-2\exp(\beta L/2)+1}{\beta^2 L^2}. \tag{4.26}$$

These equations establish bounds on the spatial aggregation error for accessibility measures that use an exponential distance decay.

In comparison with the error bounds for minimum distance measures, the equations defining error bounds for gravity-type accessibility measures with exponential distance decay have a more complicated form. In fact, the bound for square sub-areas with straight-line distances, (4.24), did not have a closed-form expression.

		Circular Sub-Areas		Square Sub-Areas			
				Straight-line		Manhattan	
β	L	Lower	Upper	Lower	Upper	Lower	Upper
.3	.25	.9753	1.025	.9718	1.029	.9633	1.038
.3	.50	.9514	1.051	.9444	1.059	.9282	1.078
.3	1.0	.9054	1.106	.8924	1.123	.8623	1.164
.3	2.0	.8208	1.224	.7978	1.263	.7464	1.360
.3	4.0	.6772	1.506	.6413	1.606	.5655	1.877
.5	.25	.9593	1.043	.9535	1.049	.9397	1.069
.5	.50	.9204	1.087	.9094	1.101	.8836	1.135
.5	1.0	.8480	1.183	.8280	1.214	.7829	1.291
.5	2.0	.7216	1.405	.6891	1.481	.6193	1.683
.5	4.0	.5285	2.000	.4850	2.235	.3996	2.952
1.0	.25	.9204	1.087	.9094	1.101	.8836	1.135
1.0	.50	.8480	1.183	.8280	1.214	.7829	1.291
1.0	1.0	.7216	1.405	.6891	1.481	.6193	1.683
1.0	2.0	.5285	2.000	.4850	2.235	.3996	2.952
1.0	4.0	.2970	4.195	.2562	5.373	.1869	10.21

Table 4.2: Table of aggregation bounds for gravity-type accessibility measures. The values for circular sub-areas and square sub-areas using Manhattan distances were calculated using equations (4.23) and (4.26) respectively. The values for square sub-areas using straight-line distances were found through the use of numerical integration on equation (4.24).

In order to examine how worst-case aggregation error varies by the size of the sub-area, L, and the value of the distance decay parameter, β, Table 4.2 tabulates the bounds on the aggregation error for $L = 0.25, 0.5, 1, 2, 4$ and $\beta = .3, .5, 1$. The range of the error bounds becomes larger for both larger values of L and β. The bounds for circular areas are tighter than those for square sub-areas. Furthermore, for square sub-areas, the bounds for straight-line distances are, as expected, tighter than those using Manhattan distances. For small sub-areas, the error bounds are fairly tight. For example, with $L = 0.25$ and $\beta = 0.5$, the true value of accessibility is within 5% of the calculated value for square sub-areas and straight-line distances. Nevertheless, these values grow rapidly with larger sub-areas. If the sub-areas were, instead, of length $L = 4$, then the error bounds on the accessibility are approximately -50% to +125%. These error bounds essentially state that the calculated aggregate accessibility of this sub-area could be only loosely coupled with the actual accessibility of individuals living in this area.

An empirical example of this, using the data set examined in Chapter 7, can be

evaluated for census *distritos* in the Central Valley of Costa Rica. The median area of the distritos for the 1984 (most recently published) census was 8.86 km^2. Consider the aggregation error bounds using the distrito of median area roughly square so that each side is approximately 2.98 km. Assuming straight-line distances and an exponential distance deterrence of $\beta = 0.5$, then numerical integration of equation (4.24) yields the following estimates for the error bound

$$0.5790 \leq F^A \leq 1.806,$$

or between -40% and +80%. These large error bounds could make the use of potential demand for health care aggregated at this level highly problematic in evaluating accessibility using a gravity-type measure. Moreover, since the size and shape of each *distrito* is different, the aggregation error bounds for each *distrito* are different. Thus, similar to what Fotheringham *et al.* [1995] note in the context of facility location models, the observed pattern of accessibility could indeed be an artifact of the manner in which the population is aggregated, rather than capturing the actual underlying accessibility of the population. These error bounds can be greatly reduced through the use of smaller population sub-areas. Using 500 metre square grid cells rather than irregularly shaped polygons, the corresponding error bounds are approximately $\pm 10\%$ which is a significant reduction in the potential level of error. Therefore, the size of the sub-areas that are being used is critical to the level of aggregation error in measuring accessibility. As illustrated above, the use of irregularly-shaped regions (such as *distritos*) with conventional modelling approaches results in potentially large aggregation errors.

4.3 Representing Populations with a Grid

The previous section established that the size of the geographic sub-areas has an important effect on the level of spatial aggregation error present in distance measures when evaluating accessibility. Moreover, the use of irregularly-shaped administrative zones or districts as demand sub-areas causes two types of difficulties. First, these zones are often quite large and can create substantial aggregation errors. Second, since their size and shape varies, the magnitude of the aggregation error can vary dramatically over space. This makes it difficult, if not impossible, to make meaningful comparisons of accessibility between zones.

In recognition of these problems, rather than using zone- or polygon-based methods for representing populations, there has been recent interest in grid-based representations of populations [Bracken, 1993]. This approach has several important advantages over conventional polygon-based representations [Martin, 1989]. Unpopulated regions remain unpopulated in a grid-based representation, which avoids the problem of calculating accessibility for uninhabited areas. Grid-cells also avoid the problem of representing large and diverse populations by a single polygon centroid. Moreover, grid-cells are of uniform size and shape so that the effects of aggregation are similar. Thus, more meaningful comparisons of accessibility can be made. Furthermore, the equations presented in the previous section allow for the calculation of error bounds on the level of spatial aggregation error.

Recently, there have been several studies of accessibility that use grid-based representations of populations. Martin and Williams [1992] used a 100 metre population grid of the city of Bristol in the U.K. for evaluating the accessibility of primary health care. Similarly, Rosero [1993] used a two kilometre grid to evaluate the physical accessibility to health facilities in Costa Rica and Geertman and Van Eck [1995] integrated a grid-based accessibility calculation into a Geographic Information System (GIS) to calculate accessibility using road network distances in the Netherlands.

4.3.1 Disaggregating Populations to Grid Cells

Although there are numerous advantages to using a grid-based population representation for evaluating accessibility, population data are typically more often available either as a set of points or polygons. Thus, in order to represent a population with a grid, it is necessary to convert the population counts from their original area- or point-based representation to a grid representation.

Many different techniques have been proposed for the interpolation of both point and area data [Lam, 1983]. One category of interpolation techniques consists of moving average methods [Ripley, 1981, Ch. 4]. For a particular grid cell under consideration, these methods select a set of neighbouring data points, typically within a "window" around the grid cell, and sets the interpolated value for the grid cell as a weighted average of these observations where the weighting factor increases the nearer the data point is to the grid cell [Ripley, 1981]. However, the interpolated value is highly dependent upon the method used to select the neighbouring observations and the type

of weighting function applied [Gold, 1989]. Further, these methods are easily affected by an uneven distribution of data points [Lam, 1983]. Some of these difficulties can be addressed using the geostatistical technique of kriging, which regards the surface to be interpolated as a regionalized variable with a certain degree of continuity [Lam, 1983]. In this method the weights are chosen so as to be unbiased and minimize the estimation variance [Oliver and Webster, 1990]. Nevertheless, difficulties exist with the selection of the appropriate neighbourhood size and the appropriate model form [Lam, 1983].

Another class of interpolation methods use tesselations and triangulations of the original data set and involve calculating a Voronoi diagram or a Delauney triangulation[4] One tesselation-based interpolation method, known as natural neighbour interpolation, was proposed by Sibson [1981]. To determine the interpolated value of a given sample point using this methods, the point is added to a Voronoi diagram of the existing data points. The data points that are neighbours to the added interpolation point are used to calculate the interpolated values. In the simplest case, these data points are then weighted by the proportion of the area of the Voronoi polygon associated with the newly added sample point that was originally contained within their respective Voronoi polygons[5]. Compared to the moving average methods, the natural neighbour method has the advantage of not requiring the specification of either a weighting function or a "window" size and automatically adjusts the interpolation parameters to the set of existing data points. Both Sibson [1981] and Gold [1989] reports good results with the application of this technique to elevation data.

Despite these advantages, Bracken and Martin [1989] note that "established interpolation techniques for point and area data are inappropriate in the case of [population] centroids" (p. 539) and, in a series of recent papers, they develop a method of disaggregating population counts from a set of census centroids to a grid using a variable-kernel density estimator [Martin, 1989; Bracken and Martin, 1989; Martin and Bracken, 1991; Bracken, 1993; Bracken and Martin, 1995]. As opposed to the previously discussed methods, the Bracken and Martin method, discussed in further detail subsequently, has the advantage of using population counts directly and ensuring that the total population of the grid equals the population of the centroids. It should be noted that this

[4]A Voronoi diagram partitions space into a series of Voronoi polygons. For each data point, there is an associated Voronoi polygon which consists of the region that is nearer to the generating central data point than any other data point. A Delauney triangulation can be formed by joining those data points whose regions share an edge. For a complete treatment of Voronoi diagrams, see Okabe *et al.* [1992].

[5]Gold [1989] provides a clear illustration of this procedure.

method has the disadvantage of requiring the specification of both a maximum window size and an appropriate distance decay function. However, Bracken and Martin [1989] note that the grid cell population counts are "much more critically controlled by the analysis of the centroid locations" and that, consequently, "the distance-decay concept becomes a secondary matter" (p. 540). Although it would be possible to adapt other interpolation methods to the problem of disaggregating population counts, it was decided to use the Bracken and Martin method, which was specifically designed for this problem, as a basis for the disaggregation process.

This method operates as follow. Given a set, $\mathcal{N}_P = \{1, \ldots, N_P\}$, of population points with an associated population P_j, $j \in \mathcal{N}_P$, the Bracken and Martin method estimates the population of a grid cell $\widehat{P}_i$, $i \in \mathcal{N}_G$, where $\mathcal{N}_G = \{1, \ldots, N_G\}$ is the set of grid cells. The basic algorithm places a "window" of a given size over each population point. The size of this window is adjusted according to the density of population points around the point so that the window is smaller in more densely settled areas. The grid cells falling within this window are then assigned a weighting representing their share of the population of that point. Grid cells falling in areas that are unpopulated, *e.g.* water, forests, or remote mountainous areas, are masked so that their weighting is zero. The weightings are then re-scaled so that the total weight for a given population point is one. Finally, the grid cells are assigned population according to their weight. This procedure is described further below.

1. The window associated with each point is initially defined as a circle of radius R and is then adjusted to the average distance from point j to the other population points falling within the radius. Thus, if $\mathcal{R}_j$ is the set of other population points within the radius,

$$\mathcal{R}_j = \{k | k \in \mathcal{N}_P, 0 < D_{jk} \leq R\} \tag{4.27}$$

then the radius of the adjusted window is

$$R_j \leftarrow \begin{cases} \frac{1}{|\mathcal{R}_j|} \sum_{k \in \mathcal{R}_j} D_{kj} & |\mathcal{R}_j| > 0 \\ R & |\mathcal{R}_j| = 0 \end{cases} \tag{4.28}$$

where $|\mathcal{R}_j|$ is the number of population points within distance R of point j.

2. The initial weighting of cell i with respect to point j is calculated as

$$W'_{ij} = \begin{cases} \rho_i \left(\dfrac{R_j^2 - D_{ij}^2}{R_j^2 + D_{ij}^2} \right)^{\alpha} & D_{ij} \leq R_j \\ 0 & \text{otherwise} \end{cases} \tag{4.29}$$

where α is a parameter representing the steepness of the decay in the weighting function and ρ_i represents the propensity of a grid cell to be populated and is defined as

$$\rho_i = \begin{cases} 1 & \text{if the grid cell could potentially be populated,} \\ 0 & \text{if the grid cell is known } \textit{a priori} \text{ to be unpopulated.} \end{cases} \tag{4.30}$$

3. The re-scaled weights are then calculated as

$$W_{ij} = W'_{ij} / \textstyle\sum_{j \in \mathcal{N}_P} W'_{ij}. \tag{4.31}$$

These new weights are defined so as to preserve the total population assigned within the window.

4. Finally, the population is assigned to each grid cell $i \in \mathcal{N}_G$ by

$$\widehat{P}_i = \textstyle\sum_{j \in \mathcal{N}_P} W_{ij} P_j \tag{4.32}$$

where $\widehat{P}_i$ is the estimated population of grid cell i.

This technique was used to develop a population grid for the United Kingdom using the 1981 census [Bracken, 1993] and later extended to the 1991 census [Bracken and Martin, 1995]. The population points used were the approximately 130 000 centroids of census enumeration districts. The population from these points was disaggregated to a 200 metre grid database. The disaggregation resulted in approximately 330 000 populated grid cells out of the 16 million cells in the national grid.

Bracken [1993; 1995] suggests many possible uses for this grid database. These uses include modelling population change over time, evaluating policy options by comparing the distribution of resources to a surface of "need," and mapping the spatial dis-

tribution of socio-economic indicators such as unemployment levels, indices of overcrowding, and the existence of sub-standard dwellings.

4.3.2 Extensions to the Bracken and Martin method

The algorithm described in the previous section only utilizes two pieces of information: the spatial distribution of the population points, and whether a grid cell can be populated or not. Often, however, further information is available about the population in the study area. This section describes two extensions to the Bracken and Martin method that incorporate additional information.

Grid Cell Classification

The Bracken and Martin method only considers two classes of grid cells. A grid cell is either known *a priori* to be unpopulated or it can be potentially populated. However, it is a simple extension to this method to incorporate any available information on the existing settlement pattern. For example, suppose that grid cells are classified into three different land use classes: urban, rural, and unpopulated. Typically, it would be expected that urban areas are more densely populated than rural areas. This information could be incorporated into the disaggregation method by allowing the propensity of a grid cell to be populated, ρ_i, to take on values other than zero or one.

Thus, for each grid cell i, one could define the propensity, ρ_i, to be

$$\rho_i = \begin{cases} \rho_U & \text{if cell } i \text{ is classified as urban,} \\ \rho_R & \text{if cell } i \text{ is classified as rural,} \\ 0 & \text{if cell } i \text{ is unpopulated.} \end{cases} \tag{4.33}$$

where $\rho_U, \rho_R > 0$. Note that step 3 of the procedure re-scales the weightings to ensure that the total population is unchanged. Different values of ρ_U and ρ_R only affect the allocation of population between various grid cells and do not affect the total allocated population. Thus these parameters may assume any non-negative real numbers.

In order to use this method, values of ρ_U and ρ_R are required. One good estimate for these values would be the average population density within each grid cell classification. Often, the population density is available from published sources, from other

ancillary sources of data, or it can be estimated from known housing densities and occupancy ratios. However, in the absence of such ancillary data, it is possible to estimate the population density from the existing data. This is done by classifying each population point by the class of the grid cell that it lies within. Thus, each population point that lies within an urban area would be classified as urban. Then population density estimates for each class could be obtained by dividing the total population for each class by its total area. These values can be used as the ρ_U and ρ_R values. Of course, it is possible to increase the number of classifications of the grid cells beyond the three discussed here through a similar procedure.

Area Population Data

In addition to points, population data are also often available in the form of population counts for irregularly-shaped administrative polygons or districts. Moreover, population counts are often available at different levels of aggregation with census areas typically available at the aggregate level and points available for smaller sub-census area units. Unfortunately, the Bracken and Martin method does not utilize the additional spatial information available from these area features.

The easiest way to utilize area population data would be to choose the centroid of each area as the population point and apply the Bracken and Martin method to a grid of given spatial resolution. Unfortunately, the application of this procedure does not ensure that the known population of an area balances with the total disaggregated population of all the grid cells within that area. However, it is possible to extend the Bracken and Martin method to incorporate this information. This extension adjusts the estimated grid cell population counts *a posteriori* to reflect the known population of each area. An additional complication is that areas that have no grid cells associated with them must be considered separately. The population of these areas are transferred directly to the population grid and are not included in the Bracken and Martin method. This procedure is described below.

1. Suppose that a study region is partitioned into N_A areas. If there is not a separate data set of population points, then define the population points as the centroids of the areas. Otherwise use the existing population points. For each area k, determine $\mathcal{P}_k$, the set of population points contained in that area, and $\mathcal{G}_k$, the set of grid cells whose centroids are located within area k. Note that it is possible that

several areas may not have any grid cells associated with them because the size of the areas is small compared to the grid cells. For example, in a densely populated urban region the census areas are often small. Since these areas do not have grid cells associated with them, the population cannot be adjusted for these areas. Instead, the populations associated with these areas are not disaggregated to the grid cells using the Bracken and Martin method and are, instead, considered later. Thus, define the subset of population points to be disaggregated as

$$\mathcal{N}_{\mathcal{P}}{}' = \bigcup_{\substack{1 \le k \le N_A \\ \mathcal{G}_k \neq \emptyset}} \mathcal{P}_k. \tag{4.34}$$

Define the subset of population points that are in areas not associated with grid cell as

$$\overline{\mathcal{N}}_{\mathcal{P}} = \mathcal{N}_{\mathcal{P}} - \mathcal{N}_{\mathcal{P}}{}'. \tag{4.35}$$

2. The population points contained in areas with no associated grid cells, $\overline{\mathcal{N}}_{\mathcal{P}}$, need to be allocated to the population grid. Since these points are located in areas that are typically small in comparison to the grid cells, the population counts for these points are directly transferred to the grid cells without use of the window disaggregation method. Thus, for each grid cell i, define $\widetilde{P}_i$ as the sum of the population of the points located within that grid cell and in set $\overline{\mathcal{N}}_{\mathcal{P}}$.
3. Use the Bracken and Martin method to disaggregate the subset of the population points, $\mathcal{N}_{\mathcal{P}}{}'$, to calculate $\widetilde{P}_i'$, the initial estimate of the population of grid cell i.
4. For each area k with grid cells associated with it, calculate C_k, the ratio of the actual population of that area to the estimated population of the area

$$C_k = \frac{\sum_{j \in \mathcal{P}_k} P_j}{\sum_{\ell \in \mathcal{G}_k} \widetilde{P}_\ell'} \qquad \mathcal{G}_k \neq \emptyset. \tag{4.36}$$

Thus C_k is a weighting factor for area k that adjusts the population of the grid cells in area k to be the known population of the area. Note that adjusting the grid cell population estimates by C_k does not affect the overall population.

5. Finally, the two grid cell population estimates are combined, $\widetilde{P}_i$ from the population points in $\overline{\mathcal{N}}_{\mathcal{P}}$, and $\widetilde{P}_i'$ from the population points in $\mathcal{N}_{\mathcal{P}}{}'$. The latter popu-

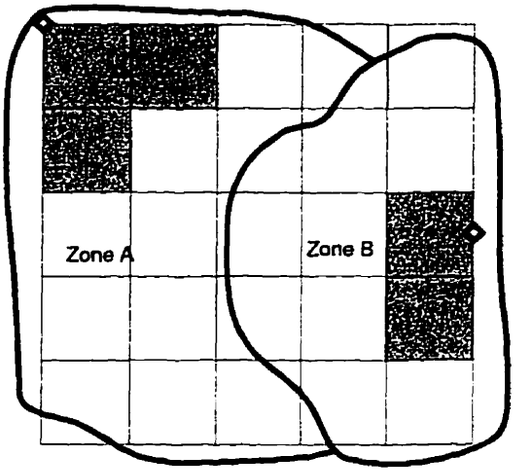


Figure 4.4: Example of disaggregating population to a grid. The ◇ indicates aggregate population points while the grey cells indicate urban areas.

lation figures are adjusted by the weighting factor C_k. Thus the final population estimate for a grid cell i in an area k is

$$\widehat{P}_i = \widetilde{P}_i + C_k \widetilde{P}'_i. \tag{4.37}$$

This process is best illustrated with a simple example. Consider an area made up of two population zones shown in Figure 4.4. The total population of Zone A is 7400 and the population of Zone B is 4700. The population point associated with each zone is marked with a ◇ on the diagram. This example disaggregates the population associated with these two zones onto the five kilometre by five kilometre grid. Thus, each grid cell is defined by a one kilometre by one kilometre square. For convenience, the grid cells are numbered in a row-major order so that cells 11, 15, 51, and 55 are the top left, top right, bottom left, and bottom right cells respectively. Furthermore, there are five grid cells classified as urban marked on the grid while all the other grid cells are rural. For the purposes of this example, the propensity of urban cells to be populated is arbitrarily set to 1.5 while the rural propensity is taken to be one. Finally α is defined to be 1.

From the diagram, the grid cells can be partitioned into two sets depending on the

zone in which its centroid lies. Thus, the cells in Zone A, $\mathcal{G}_A$ and Zone B, $\mathcal{G}_B$, are

$$\mathcal{G}_A = \{11, 12, 13, 14, 21, 22, 23, 31, 32, 41, 42, 51, 52, 53\}$$
$$\mathcal{G}_B = \{15, 24, 25, 33, 34, 35, 43, 44, 45, 54, 55\}.$$

For this particular example both zones have grid cells associated with them.

The distance between the population points is 5.59 km. Thus, this value is the radius of the adjusted window around each population point. Consider the weight of cell 11 for population point A. The distance $D_{11,A}$ is 0.707km and $\rho_{11} = 1.5$. By applying formula (4.29), the weight is calculated to be $W'_{11,A} = 1.4528$. This process is repeated for all the grid cells for population point A and the sum of these weights is 10.94. Thus, the re-scaled weighting of cell 11 with respect to population point A is 0.1328. Equally, the weighting of cell 11 with respect to population point B is 0.0122. Therefore, the initial population estimate rounded to the nearest integer of cell 11 is

$$\widehat{P}_{11} = 0.1328 \times 7400 + 0.0122 \times 4700 = 1053.$$

The initial population estimates for the grid cells, shown in Table 4.3, are obtained by repeating this process for each grid cell.

	Initial Estimates					Adjusted Estimates				
Col.	Row 1	Row 2	Row 3	Row 4	Row 5	Row 1	Row 2	Row 3	Row 4	Row 5
1	1053	1022	606	501	379	1006	976	579	478	409
2	964	638	586	503	399	920	609	560	544	431
3	515	528	501	445	544	492	504	541	481	588
4	353	377	370	336	431	337	360	399	362	466
5	181	210	214	208	237	173	200	204	224	256

Table 4.3: Initial and adjusted grid cell population estimates. The population estimates do not total correctly due to rounding.

However, for each zone, the sum of the initial population estimates within each zone are not equal to the total population of the zone. For example, for population Zone A, the total population of the associated grid cells, $\mathcal{G}_A$, is equal to 7748. The initial population estimates are, therefore, multiplied by a weighting factor to ensure that the

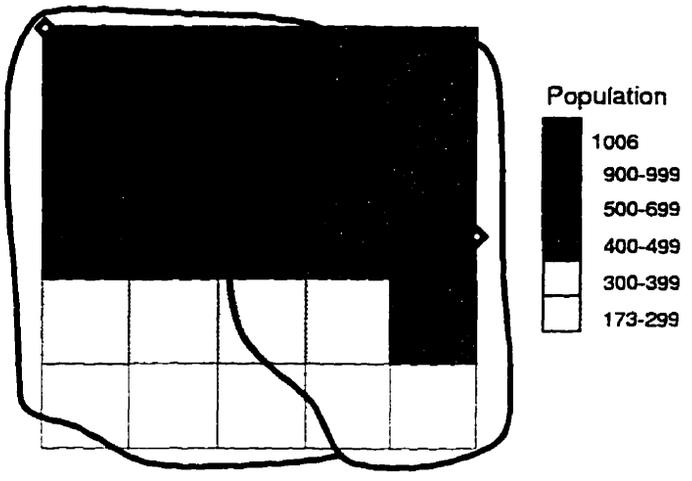


Figure 4.5: Adjusted population estimates for disaggregation example.

population of grid cells within a zone is equal to the zone population. For example, the weighting factor for Zone A is

$$C_A = 7400/7748 = 0.955.$$

Thus, the adjusted population estimates, shown in Table 4.3, are obtained by multiplying the grid cells by the weighting factor of their corresponding zone. These population estimates are also shown in Figure 4.5.

4.4 Summary

This chapter examined the effects of aggregation on the evaluation of accessibility. Although the generic model of accessibility specified in the previous chapter considers accessibility in terms of the individual, the operationalization of this model on a regional scale requires the use of aggregate-level accessibility measures.

First, methods of aggregating individuals were discussed. One method of creating an aggregate accessibility measure was to construct an average individual and use this individual's accessibility as a surrogate for the accessibility of the population. However, the aggregation error increases with the variance of the relevant characteristics and this method can mask important variations in accessibility among individuals. The aggre-

gation error can be reduced by partitioning the population into groups to minimize the within group variance of the socio-demographic characteristics of the individuals and the distances of the individuals to the facilities. This suggests that the population should be partitioned by the demographic characteristics that have an important impact on accessibility, and by space so as to reduce the variation in the distances from the individuals in that group to the facilities. For most potential measures of geographic accessibility, the effect of socio-demographic characteristics is de-emphasized and, furthermore, grouping even by a small number of these characteristics greatly reduces aggregation error.

For minimum distance measures, worst-case error bounds were derived for circular sub-areas with straight-line distances and square sub-areas with straight-line and Manhattan distances. For all of these cases, the worst-case aggregation error increased linearly with the size of the sub-area. For the gravity-type accessibility measure, the worst-case aggregation error bound was a fraction of the accessibility. Worst-case error bounds were calculated for different sub-area sizes and distance deterrence parameters. Potentially large aggregation errors could occur with the use of gravity model accessibility measures. These large error bounds make the use of these measures problematic with aggregated data.

Next, the chapter considered disaggregation of population data from relatively large zones or polygons to a grid of given spatial resolution. Two extensions to the Bracken and Martin method were presented and, using a simple hypothetical example, the process of spatial disaggregation was described. Through choosing an appropriate grid cell size, with known aggregation error bounds, the potentially adverse impacts of this problem can be controlled. Consequently, the reliability of accessibility measures is enhanced and they can be used as a basis for improving access to health care through model-based planned intervention. The means by which this can be achieved is discussed in the next chapter.

Chapter 5

A Generic Accessibility Optimization Model

This chapter builds upon the previous two chapters by first assessing possible strategies for improving the geographic accessibility of a health care system using a facility-oriented optimization approach. Criticisms of the use of optimization methods, and particularly the use of facility location models, for health care and service development planning are considered. After this, generic planning objectives, suitable for optimizing geographical accessibility to health care, are discussed. These objectives examine *both* the efficiency and equity of the distribution of accessibility in the target population. Finally, formulations are presented for the generic Accessibility Optimization Problem (AOP) and for two specific subproblems: the Facility Location Subproblem (FLS), and the Resource Allocation Subproblem (RAS).

5.1 Strategies for Improving Accessibility to Health Care Services

The need for primary health care services among poor and geographically dispersed target populations is great. In this context, Phillips [1990] notes that health care resources, particularly in developing countries, are both inequitably distributed and inefficiently allocated. Thus, it is crucial for health development planners and decision makers to make the best use of scarce resources committed to primary health care provi-

sion in order to improve equity and efficiency in the distribution of health care resources and thereby strengthen the well-being of target populations.

There are important barriers to achieving effective planning for primary health care. Health care planning in developing countries often tends to be episodic and capricious so that local system changes are often implemented without consideration of their effects. Further, health care systems are often planned in a fragmented, and sometimes conflicting, manner through funding provided by public, private, charitable, and aid sources. These problems are further confounded by the fact that not only are resources inadequate but services are also sub-optimally located [Oppong and Hodgson, 1994].

An additional difficulty relates to the complexity of the planning process itself. As mentioned above, it is important to ensure that health care is accessible to a widely dispersed and unevenly distributed population. The evaluation of health care accessibility is a complicated spatial problem and providing accessible health care in an efficient and equitable manner is a complex issue. Phillips [1990] notes that difficulties in improving health care services are not solely related to the scarcity of financial resources but also result from practical difficulties in management and planning. Thus, there is an important need to assist health development planners in devising potential strategies for improving both social accessibility, through the examination of social and institutional barriers, and geographic accessibility, which is the emphasis in this thesis.

Depending on the nature of the services being provided, there are many possible strategies available for improving geographic accessibility to these services. Mosely [1979] identifies four broad types of policy options: facilitating mobility of the population; making the service providers mobile; inducing the population to live nearer to the service providers; or taking a facility-oriented approach through modifying the characteristics of the service delivery system.

The methods developed in this chapter, consistent with the general approach adopted in the thesis, emphasize the latter approach, by seeking to change the nature of the current system of facilities to improve user accessibility. Recall that the generic model, presented in Chapter 3, conceptualizes the accessibility of a given individual to a facility as a function of the distance from the individual to the facility, the attraction of a facility, and the facility's congestion. This suggests three ways of modifying accessibility: reducing distance, increasing attraction, or decreasing congestion. However, the congestion of a facility is a function of the spatial distribution of the population with respect to a facility and, thus, can be only indirectly modified through changing

a facility's location or by increasing service capacity. Hence, the optimization models discussed in this chapter assist decision makers by determining *(a)* the locational configuration of facilities to reduce distance deterrence, and *(b)* the allocation of resources to improve attractiveness.

It should be re-iterated that these two strategies are not the only possible approaches available for improving access to health care services. For example, the low quality of services provided at existing service providers may reduce the accessibility of primary health care services [Annis, 1981]. Further, insitutional and social factors can also act as barriers to accessibility. However, these types of problems are hard to quantify, vary dramatically in importance for different areas, and generally require a high level of local knowledge. Thus, they are not amenable to being formulated within a generic mathematical framework. It is important to note, however, that once a suitable mathematical formulation is derived and tested, it can be supplemented by direct analysis of qualitative factors.

The models formulated in this chapter are generic and do not depend on local conditions for their application. Although the strategies for improving accessibility developed by these models are not appropriate for every situation, they provide very important information to health development planners in cases where additional resources are being allocated to a system or where new facility locations are being planned. In these cases, it is important to allocate resources and to locate new facilities optimally so that scarce resources are used most effectively when the system is expanded [Phillips, 1990]. Another potential application for these models is in situations where the resources are being removed or facilities are being closed. Furthermore, these models can also be used to generate an optimal pattern of facility locations and resource allocations and to compare the current system configuration to the optimal configuration. The information provided by this particular analysis is a very important tool in examining how resources within the existing health care system may be deployed with greater efficiency and equity.

Despite the potential advantages of using optimization methods to assist planners and decision makers, several authors have criticized the use of this approach. In particular, the use of facility location models for health care and service development planning in developing countries has been criticized. The next section considers these criticisms.

5.2 The Appropriateness of Using Optimization Models for Primary Health Care Planning

As discussed in Chapter 2, optimization models determine the best values of a set of one or more decision variables, such as facility locations or resource allocations, with respect to one or more planning objectives. Facility location models are a type of optimization model that examine the problem of locating new facilities with respect to existing facilities and the location of clients. These models have been used extensively in developed countries in many application areas such as locating warehouses, emergency service planning, education planning, and health care planning [Hansen *et al.*, 1987]. Facility location models optimally locate service delivery points and, therefore, can improve the efficiency and equity of the distribution of resources in the system.

Rushton [1984] strongly encourages the use of optimization models for service development planning in the developing world. He asserts that since resources, such as money, equipment, and personnel, are very limited, it is crucial that these scarce resources be used as efficiently and as fairly as possible. Furthermore, Rushton [1988] indicates that the potential use of facility location models has expanded over the years and he identifies six potential application areas:

1. determining the optimal set of locations with respect to pre-defined objectives;
2. comparing actual systems of facilities to their normative counterparts;
3. finding a set of additional facility locations to add to the existing set;
4. evaluating the benefits and costs of constraints on real-life decisions;
5. assessing the quality of recent locational decisions; and,
6. examining alternative decision-making principles by determining the system of facilities that would develop if these principles were used.

However, there are significant barriers to the adoption of facility location models for infrastructure planning, such as primary health care provision, in developing countries. Rushton [1988, p. 99] notes that "the major international organizations either explicitly reject or ignore formal facility location models in favour of more traditional, subjective, graphical approaches." Major aid organizations that have rejected this approach include the World Health Organization (WHO) [Kleczkowski and Pubouleau, 1976]

and United States Agency for International Development (USAID) [Rondinelli, 1985]. In fact, Rondinelli [1990] explicitly rejects the use of facility location models since he expects that:

1. the models would give the same solution as his proposed urban functions methodology;
2. the demand data for the facility location models are very difficult to obtain; and,
3. the models are too sophisticated to be used or understood by the relevant authorities.

Instead, Rondinelli [1985] proposes an urban functions approach. This very flexible and process-oriented approach is a ten stage process of data collection, description, analysis, planning, and monitoring [Rietveld, 1990]. However, as noted by Rushton [1993] the link between analysis and planning within the urban functions approach can often be very weak.

Rietveld [1990] answers Rondinelli's criticisms by pointing out that facility location models can be a useful technique within the urban functions approach. Furthermore, he remarks that, at the very least, facility location models can be used to verify the locational choices of the urban functions approach and that, contrary to Rondinelli's claim, the data requirements are not excessive. Moreover, facility location models are not too sophisticated to be used in developing countries. As Rushton [1993, p. 321] remarks "people in [developing countries] are as capable as their counterparts in developed countries of understanding the concepts of location-allocation models and, if properly presented, their results." Rushton further criticizes the urban functions approach for both its ambiguity and also for the fact that it allows planners "to justify virtually any pattern of infrastructural investment" (p. 319) and "to advocate any policy and cite supporting evidence."

In contrast, Phillips [1990] also questions the utility of facility location models in primary health care planning and notes that "optimally designed spatial health care delivery systems are not practical in the Third World" (p. 145) due to problems of data availability and reliability. He further notes that even if data are available, facility location models lack flexibility. He writes:

> The procedure allocated facilities to locations that will serve the most people yet minimize travel distances. However, theoretical answers do tend

> to neglect social and economic variations among populations, and thus can ignore equity considerations (p. 146).

Thus, Phillips argues that facility location models require modifications for use in a developing country context and that "such refinements may in practice be difficult to achieve" (p. 147).

Recent developments have tended to minimize the importance of these criticisms. The WHO recently stated that health information is of "crucial importance" for developing and implementing district health systems and recommends that "essential data ... must be identified, particularly as regards the problems of equity, efficiency and quality of care" [1994, p. 28]. Thus, the problem of data availability may become less important with the WHO encouraging strongly the collection of the very data needed for locational analysis. Moreover, Oppong and Hodgson [1994] also observe that critics often exaggerate the data requirements for facility location models. Finally, as discussed in Chapter 2, there are many different types of facility location models. Many of these models have objectives other than adjusting facility accessibility by minimizing distances. Although not yet extensively applied for primary health care planning in the developing world, new facility location models and techniques, such as the interaction-based model of Oppong [1992] and the models developed in this thesis, can incorporate additional factors and explicitly incorporate equity considerations.

Perhaps a more fundamental criticism of an optimization approach to service development planning is presented by Gore [1991a; 1991b], who states that optimization models incompletely relate the social aspects of service provision by giving preeminence to the spatial aspects. Specifically, Gore [1991a] criticizes the use of facility location models for three main reasons. First, these models consider only one policy option to improve geographical accessibility, namely, reducing the distance between the person and the facility, while ignoring other options, such as making the service mobile or facilitating the mobility of the person. Second, facility location models separate the facility location decision from other aspects of service provision so that the trade-off between quality of service and the number of facilities cannot be considered. Third, facility location models separate the spatial impact of changes in the location of facilities from the social and economic impact of these changes. Moreover, Gore criticizes facility location models because they do not consider whether additional supplies or greater accessibility is a benefit; these models cannot consider either true equity or efficiency

because they are only capable of considering it in a spatial sense; and they perpetuate existing patterns of inequality by considering users in an aggregate sense and using past interaction patterns so that non-users are ignored. Instead, Gore [1991b] proposes an entitlement approach to service development planning.

The points that Gore raises are interesting and highlight some important limitations of the optimization approach. First, optimization models do assume that improving access to services, such as primary health care, will accrue benefit to the users. This is a fundamental assumption of the models, although not a great limitation. Second, the models are not typically intended to capture every nuance of reality, that is why they are models and why it is important for them to be considered within the context of providing information to planners and decision makers to assist them with the decision-making process. Gore's comments underline the importance of identifying appropriate application areas for these types of models, much as Rushton [1988] does.

Contrary to Gore's criticisms, there are *no* theoretical barriers to the consideration of individual characteristics when using an optimization approach. As discussed in the previous chapter, the data collection and computational difficulty of the models do make this approach problematic. However, these difficulties are present in any decision-making process if individual-level data are used in the process. Second, regional scale planning cannot consider each user individually due to difficulties in data gathering, storage, and processing. Therefore, individuals must be aggregated in some way to make the planning process feasible. Thus, aggregate data must be used in the planning process whether this process involves optimization models or not. Similarly, it is possible to include non-users when evaluating or optimizing accessibility. In fact, the generic model of accessibility, presented in Chapter 3, explicitly incorporates the option of non-use into the choice process. Equally, there are no barriers, other than data availability, to the inclusion of quality of service indicators within the analysis.

In summary, many authors have criticized the use of optimization models, such as facility location models, for service development and health care planning in developing countries. Although these models do not address every aspect relevant to this issue, they are appropriate for certain problems and belong in the "kitbag of tools that planners should bring to their work" [Rushton, 1993, p. 322]. For example, Tewari [1992, p. 34] contends that in India the lack of such tools allows "political pressure groups ... to have a significant influence on the decision-making process," thereby allowing for resources to be deployed inefficiently and inequitably.

The next section outlines several suitable planning objectives for measuring the potential geographical accessibility of a population to health care in terms of efficiency and equity.

5.3 Objective Function Formulations

An important issue in developing optimization methods for improving accessibility is identifying the criteria used to evaluate potential strategies. Much of the criticism of using these models questions whether the correct planning objectives are being optimized. Thus, the development of appropriate objective functions for the optimization problems is crucial for the successful application of these models. Towards this end, this section outlines several generic accessibility objectives that are suitable for use in many contexts. However, for specific applied studies of particular regions and health care systems, it is possible to develop specialized planning objectives that account for specific local conditions.

For private sector optimization problems, the objectives are typically to minimize costs or maximize profits. However, when the area of application involves modelling a public good or service such as primary health care provision, different planning objectives are required. Typically, these objectives are classified into two main categories: efficiency objectives and equity objectives. These objectives are now considered, building upon the definitions provided in Chapter 1.

5.3.1 Efficiency Objectives

Efficiency objectives measure the ratio of the total aggregate level of services or benefit relative to the level of inputs required to provide the services. When used in an optimization framework, efficiency objectives preferentially allocate resources so that they have the maximum aggregate benefit. In the case of health care this would allocate the most resources to facilities in areas that have the largest target populations.

One obvious efficiency objective is the total aggregate accessibility of the population. A straightforward formulation of this efficiency objective is to maximize aggregate accessibility

$$Z_F = \sum_i P_i A_i \tag{5.1}$$

where P_i is the target population and A_i is the accessibility of sub-group i. This particular objective has been widely used with minimum-distance accessibility measures. In this instance, the goal of maximizing the aggregate accessibility is equivalent to minimizing the total weighted distance to the nearest facility. This is the standard p-median facility location problem described in Chapter 2.

For accessibility measures derived from the random utility model specified in Chapter 3, several researchers have suggested the use of an alternative measure of efficiency [Wilson *et al.*, 1981; Mayhew and Leonardi, 1982; Fotheringham and O'Kelly, 1989]. Recall that within this random utility model, the expected value of the maximum perceived attractiveness (satisfaction) of an individual in group i was defined in equation (3.39) as[1]

$$V_i^* = \ln \sum_j \exp V_{ij} \tag{5.2}$$

where V_{ij} is the perceived attractiveness of facility j to an individual in population group i. Thus, a reasonable efficiency measure would be the total aggregate satisfaction of the population defined as

$$Z_S = \sum_i P_i \ln \sum_j \exp V_{ij}. \tag{5.3}$$

If we have accessibility defined as $A_i = \sum_j \exp V_{ij}$, as is typical with gravity model accessibility measures, then the standard form of this objective involves the maximization of

$$Z_S = \sum_i P_i \ln A_i. \tag{5.4}$$

5.3.2 Equity

Equity measures attempt to evaluate the fairness, impartiality, or equality of the resource distribution and service provision relative to the distribution of the target population. Thus, equity objectives allocate resources preferentially to areas or population groups with below average accessibility in order to reduce the variation in equality of access between areas and population sub-groups.

[1]The constant term has been omitted from this equation and the scale parameter, μ, is assumed to be one to be consistent with existing models of accessibility.

In contrast to efficiency, equity in access to primary health care services can be measured in many different ways. For example, in a recent paper, Mulligan [1991] describes eight different potential equity measures in the context of facility location problems. This section presents a subset of these measures that are suitable for evaluating equity in the context of potential health care accessibility.

One of the simplest and most widely-used measures of equity is the accessibility of the least accessible area or population sub-group in the study area. This corresponds to the following objective function formulation,

$$Z_M = \min_i A_i. \tag{5.5}$$

This particular objective has been widely used with minimum-distance accessibility measures. The problem of locating p facilities so as to maximize the minimum accessibility, or equivalently minimize the maximum distance to the nearest facility, is the standard p-centre facility location problem.

Another measure of equity in the distribution of resources is the coefficient of localization, discussed in Chapter 2. The coefficient of localization measures the concentration across population sub-groups of an activity or resource of interest relative to the base magnitude [Joseph, 1982]. In examining accessibility, the coefficient of localization is expressed as

$$Z_L = \frac{1}{2}\sum_i \left| \frac{P_i A_i}{\sum_k P_k A_k} - \frac{P_i}{P_T} \right| \tag{5.6}$$

where P_T is the total population. A completely equitable distribution corresponds to a coefficient of localization of zero and this coefficient increases proportionately for increasingly inequitable distributions. Joseph [1982] stresses the importance of interpreting this index with respect to its upper bound. This upper bound occurs when the sub-group with the smallest population has a non-zero accessibility while all the other sub-groups are inaccessible and thus

$$Z_L \leq 1 - \min_i P_i/P_T. \tag{5.7}$$

The mean deviation around the arithmetic mean is a simpler method of capturing the dispersion in the accessibility measures. This measure compares the accessibility

of a given population sub-group to the average accessibility of the population and is defined as

$$Z_D = \sum_i P_i \left| A_i - \overline{A} \right| . \tag{5.8}$$

Again, a completely equitable distribution is indicated by a value of zero with higher values correspond to increasingly inequitable distributions.

Finally, a common measure of equity is given by the variance of the distribution of accessibility. Since this measure squares the difference between the accessibility of a population sub-group and the average accessibility, it tends to accentuate differences that are relatively large. Mathematically, the variance can be expressed as

$$Z_V = \sum_i P_i \left(A_i - \overline{A} \right)^2 . \tag{5.9}$$

Thus, this expression measures the variance in accessibility at the individual level. In contrast to the other equity objectives, the variance measure has the important property that it is smooth, *i.e.*, the first partial derivatives of this objective are continuous with respect to changes in individual accessibility levels. This property has important advantages for use within an optimization methodology.

Thus, this section proposed to broad classes of optimization objectives. Efficiency objectives maximize the total accessibility or benefit while equity objectives attempt to distribute of accessibility or resource within the target population fairly. The next section formulates the generic Accessibility Optimization Problem (AOP) that can be used to optimize these planning objectives so as to improve the geographical accessibility of a population to health care.

5.4 The Accessibility Optimization Problem

The goal of the Accessibility Optimization Problem (AOP) is to assist health planners in developing countries to make constructive changes to the health care system in order to improve its efficiency and equity, as expressed by the planning objectives described in the previous section. As noted earlier, this problem considers two facility-oriented strategies for changing the geographical accessibility of a population to a health care

system, namely, by changing where facilities are located and by modifying the allocation of resources to the facilities.

A solution to the AOP can assist decision makers by determining a new locational configuration of facilities and allocation of resources that are optimal in terms of one or more planning objectives. Furthermore, it is possible to use the AOP in a variety of specific planning scenarios. These scenarios include adding new facilities or resources, removing existing facilities or resources, modifying the locational configuration or resource allocations, and comparing the optimal solution to the existing system. The information given by these analyses assist the decision-making of health planners by quantifying the impact of possible system changes. Moreover, these optimization strategies help promote the use of the planning process to invoke change within the system. In this context, it must be stressed that the optimized solution is strictly advisory and suggests a possible strategy as optimal, subject to a set of planning objectives and data constraints.

5.4.1 Problem Formulation

Consider a system with $N_E \geq 0$ existing facility sites, and $N_F \geq 0$ potential facility sites so that there is a total of $N_D = N_E + N_F \geq 1$ candidate facility sites. Each candidate facility site has an allocation of $N_R \geq 0$ different resource levels that impact on the attractiveness of a facility located at that site. Resources can include the number of trained personnel, the total hours of service available, and the level and type of supplies. The goal of AOP is to determine the locational configuration and the allocation of resources to new and existing facilities to optimize a vector

$$\mathbf{Z} = [Z_1, Z_2, \ldots Z_{N_Z}]$$

of $N_Z \geq 1$ planning objectives or criteria.

The first strategy for improving accessibility involves determining the locational configuration of the system, *i.e.*, whether or not a facility is located at a given candidate facility site. This locational configuration can be represented as a vector of decision variables indicating the siting decision for each candidate site,

$$\mathbf{y} = [y_1, y_2, \ldots y_{N_D}]$$

where the variable y_j, $1 \leq j \leq N_D$, is defined as

$$y_j = \begin{cases} 1 & \text{if a facility is located at site } j, \\ 0 & \text{otherwise.} \end{cases}$$

Typically, a number of sites correspond to existing facility locations, *i.e.*, y_j is constrained to be 1. These constraints on the solution can be incorporated into the optimization model by defining an additional constant binary vector

$$\mathbf{Y} = [Y_1, Y_2, \ldots, Y_{N_D}]$$

termed the *required locations vector*, whose elements are set as follows

$$Y_j = \begin{cases} 1 & \text{if a facility must be located at site } j, \textit{ i.e.}, y_j = 1, \\ 0 & \text{the siting decision at } j \text{ is determined by the model,} \end{cases}$$

for $1 \leq j \leq N_D$, and imposing the condition $y_j \geq Y_j$, $1 \leq j \leq N_D$.

The second set of decision variables concerns the allocation of the N_R resources among the facilities. To describe these allocations define the vector of allocations

$$\mathbf{s} = [s_{11}, s_{12}, \ldots, s_{N_D,N_R}]$$

where s_{jk} is the level of resource k allocated to the facility at site j. Furthermore, define Q_k to be the total level of resource k available for allocation among the facilities, and let $S_{jk}^{MIN} \geq 0$ and $S_{jk}^{MAX} \leq Q_k$ be the minimum and maximum possible allocations of resource k to the facility at site j.

Adopting this notation, the AOP can be formulated as follows.

$$\underset{\mathbf{s},\mathbf{y}}{\text{Maximize}}\ \mathbf{Z} = (Z_1, Z_2, \ldots Z_{N_Z}) \tag{5.10}$$

subject to

$$y_j S_{jk}^{MAX} - s_{jk} \geq 0 \qquad j = 1, \ldots, N_D \qquad k = 1, \ldots, N_R \tag{5.11}$$

$$s_{jk} - y_j S_{jk}^{MIN} \geq 0 \qquad j = 1, \ldots, N_D \qquad k = 1, \ldots, N_R \tag{5.12}$$

$$\sum_{j=1}^{N_D} s_{jk} \leq Q_k \qquad k = 1, \ldots, N_R \tag{5.13}$$

$$y_j \geq Y_j \qquad j = 1, \ldots, N_D \tag{5.14}$$

$$y_j \in \{0,1\} \qquad j = 1, \ldots, N_D. \tag{5.15}$$

In this formulation, constraint sets (5.11) and (5.12) ensure that the resource allocations are within the acceptable bounds if a facility is located at site j or are zero if $y_j = 0$. Constraint set (5.13) ensures that the total level of resource k allocated does not exceed the supply. Constraint set (5.14) ensures that service providers remain located at sites with existing facilities. Finally, constraint (5.15) is the integrality restriction so that $\mathbf{y}$ is a binary vector.

The AOP simultaneously determines the optimal facility locations and the optimal resource allocations. The specific form of the objective functions for the AOP are dependent upon the planning objectives and the accessibility measure selected. However, most of the planning objectives, outlined in the previous section, are non-linear so that this problem is typically a non-linear programming problem. Moreover, the AOP is a type of facility location model (discussed in Section 2.3). These models are often difficult to solve and "even some of the most basic models are computationally intractable for all but the smallest problem instances" [Church *et al.*, 1993, p. 1]. For instance, even a simple facility location model with a linear objective, such as the p-median problem, is *NP*-hard. Thus it can often be computationally infeasible to solve directly the AOP as formulated above.

One way of overcoming these potential computational difficulties is to consider two related subproblems that only optimize a single vector of decision variables at a time, *i.e.*, only consider either modifying the locational configuration of the facilities or the allocation of the resources to facilities in the system. Moreover, in many situations, decision makers and planners may only be interested in examining the impact, on accessibility, of changing either the facility locations or the resource allocations indepen-

dently of each other. To facilitate these approaches, the AOP can be partitioned into two subproblems: the Facility Location Subproblem (FLS), and the Resource Allocation Subproblem (RAS). These two subproblems can also be applied to specific planning scenarios such as expanding, contracting, and modifying the existing system as well as comparing the existing system to an optimal one. The next two sections provide generic formulations for these two subproblems.

5.4.2 The Facility Location Subproblem

The Facility Location Subproblem (FLS) modifies accessibility through changing the locational configuration of the facilities. This subproblem assumes that the resource allocations to existing facilities or potential facility sites are fixed. The allocation of resource k to site j is fixed to a constant, S_{jk}, if a facility is located at site j (*i.e.*, if $y_j = 1$). By setting $S_{jk}^{MIN} = S_{jk}^{MAX} = S_{jk}$ we have as a consequence of (5.11) and (5.12),

$$s_{jk} = y_j S_{jk}.$$

Substituting this into constraint (5.13) yields the following equivalent constraint

$$\sum_{j=1}^{N_D} S_{jk} y_j \leq Q_k.$$

With these modifications, the AOP takes the form of the following discrete FLS.

$$\underset{\mathbf{y}}{\text{Maximize}} \ \mathbf{Z} \tag{5.16}$$

subject to

$$\sum_{j=1}^{N_D} S_{jk} y_j \leq Q_k \qquad k = 1, \ldots, N_R \tag{5.17}$$

$$y_j \geq Y_j \qquad j = 1, \ldots, N_D \tag{5.18}$$

$$y_j \in \{0,1\} \qquad j = 1, \ldots, N_D. \tag{5.19}$$

Many of the facility location models discussed in Chapter 2 are special cases of the FLS, with the appropriate selection of accessibility measures and objectives. Thus, many of the solution techniques for facility location models can be applied, with suitable modifications, to the FLS. Specific examples of this are provided in Chapter 6. The FLS can be applied to the four planning scenarios outlined previously, namely, locating new facilities, closing existing facilities, moving facilities, and determining an optimal

configuration of facilities.

Two additional definitions simplify the formulation of the FLS corresponding to these four scenarios. First, it is possible to partition the binary decision vector into two subvectors so that

$$\mathbf{y} = \left[\mathbf{y}^E, \mathbf{y}^F\right]$$

where $\mathbf{y}^E$ and $\mathbf{y}^F$ are the binary vectors of siting decisions for the existing facilities and the potential facility sites respectively. Partitioning the required locations vector in the same way yields

$$\mathbf{Y} = \left[\mathbf{Y}^E, \mathbf{Y}^F\right]$$

where $\mathbf{Y}^E$ and $\mathbf{Y}^F$ correspond to the required locations vector for the existing facility and potential facility sites respectively. The FLS formulations corresponding to the four planning scenarios are outlined below.

Locating New Facilities The facility location subproblem can be used to determine the optimal locations for N new facilities. For this situation, it is assumed that no facilities are being closed so that each existing facility is a required location, and therefore the required locations vector for the existing facilities is $\mathbf{Y}^E = \mathbf{1}$, and thus, $\mathbf{y}^E = \mathbf{1}$. In addition, since no constraint is placed on siting a facility at a potential facility site, $\mathbf{Y}^F = \mathbf{0}$. Assuming no other resource constraints, this optimization problem is defined as follows.

$$\underset{\mathbf{y}^F}{\text{Maximize}}\ \mathbf{Z} \tag{5.20}$$

subject to

$$\sum_{j=1}^{N_F} y_j^F = N \tag{5.21}$$

$$y_j^E = 1 \qquad j = 1, \ldots, N_E \tag{5.22}$$

$$y_j^F \in \{0,1\} \qquad j = 1, \ldots, N_F. \tag{5.23}$$

Closing Existing Facilities The FLS can be used to determine which facilities should be closed so as to have the least negative impact on either the total accessibility, or the efficiency and equity of the system. Since the system is contracting, there are no potential facility sites so that $N_F = 0$. If all existing facilities are candidates for closure then $\mathbf{Y}^E = \mathbf{0}$. However, if one or more of these facilities are required to remain open then the appropriate entries in the required locations vector, $\mathbf{Y}^E$,

should be set to 1. In either case, the problem can be written as

$$\underset{\mathbf{y}^E}{\text{Maximize}}\ \mathbf{Z} \tag{5.24}$$

subject to

$$\sum_{j=1}^{N_E} y_j^E = N_E - N \tag{5.25}$$

$$y_j^E \geq Y_j^E \qquad j = 1, \ldots, N_E \tag{5.26}$$

$$y_j^E \in \{0,1\} \qquad j = 1, \ldots, N_E \tag{5.27}$$

where N designates the number of existing facilities to be closed.

Moving Facilities Another use for the facility location subproblem is to determine the effects on accessibility of allowing up to N facilities to move to their optimal locations. This problem can be formulated as follows.

$$\underset{\mathbf{y}}{\text{Maximize}}\ \mathbf{Z} \tag{5.28}$$

subject to

$$\sum_{j=1}^{N_E} y_j^E \geq N_E - N \tag{5.29}$$

$$\sum_{j=1}^{N_F} y_j^F \leq N \tag{5.30}$$

$$\sum_{j=1}^{N_D} y_j = N_E \tag{5.31}$$

$$y_j^E \geq Y_j^E \qquad j = 1, \ldots, N_E \tag{5.32}$$

$$y_j \in \{0,1\} \qquad j = 1, \ldots, N_D. \tag{5.33}$$

Constraint (5.29) ensures that no more that N facilities are closed, constraint (5.30) allows up to N new facilities to be located, and (5.31) makes sure that the total number of facilities remains unchanged. Thus, this strategy is equivalent to simultaneously locating up to N new facilities and closing the same number.

Determining the Optimal Facility Configuration A final use for the facility location subproblem is to find the optimal solution to the problem of locating N_E facilities, assuming that there are no existing facilities. The optimal system of facility locations can then be compared to the existing configuration. For this situation, there are no required locations and, consequently, $\mathbf{Y} = \mathbf{0}$. Thus, the facility loca-

tion subproblem can be formulated as the following optimization problem.

$$\underset{\mathbf{y}}{\text{Maximize}} \ \mathbf{Z} \tag{5.34}$$

subject to

$$\sum_{j=1}^{N_D} y_j = N_E \tag{5.35}$$

$$y_j \in \{0,1\} \quad j = 1,\ldots,N_D. \tag{5.36}$$

Note that this formulation locates N_E facilities out of the set of candidate facility sites without regard to whether a facility exists at that site or not.

5.4.3 The Resource Allocation Subproblem

Another method of changing existing levels of accessibility involves modifying the allocation of resources to the various facilities in a health care system. Assuming that no facilities are opened or closed greatly simplifies the formulation of the AOP. For this situation, $Y_j = 1$ for all sites and, therefore, $y_j = 1$ as a consequence of constraint (5.14). Thus, the AOP can be reformulated as the Resource Allocation Subproblem (RAS) as follows.

$$\underset{\mathbf{s}}{\text{Maximize}} \ \mathbf{Z} \tag{5.37}$$

subject to

$$S_{jk}^{MIN} \leq s_{jk} \leq S_{jk}^{MAX} \quad j = 1,\ldots,N_D \quad k = 1,\ldots,N_R \tag{5.38}$$

$$\sum_{j=1}^{N_D} s_{jk} \leq Q_k \quad k = 1,\ldots,N_R. \tag{5.39}$$

The RAS lacks the combinatorial component of the AOP. Different solution methods are possible depending on the specific structure of the accessibility measure and the objectives. Typically the objective functions would be non-linear functions of the $N_D \times N_R$ variables s_{jk} that are subjected to $N_D \times N_R$ simple bound constraints (5.38) and N_R linear constraints (5.39).

In addition, the RAS can also be applied to the four planning scenarios outlined previously. For this problem, these four scenarios correspond to: allocating new resources, removing existing resources, re-allocating resources among facilities, and comparing the current resource allocations to an optimal allocation. The application of the RAS to these planning scenarios does not affect the problem formulation given by (5.37) through (5.39). Instead, these four scenarios specify the values of the minimum and

maximum facility resource levels, S_{jk}^{MIN} and S_{jk}^{MAX}, and the total resource levels, Q_k, in terms of the existing system configuration in addition to scenario dependent parameters. The existing configuration can be described by Q_k^{CUR} and S_{jk}^{CUR}, where Q_k^{CUR} represents the total level of resource k currently available in the system and S_{jk}^{CUR} represents the existing level of resource j at service provider k. Outlined below are descriptions of how the four planning scenarios affect the minimum and maximum facility resource bounds (S_{jk}^{MIN} and S_{jk}^{MAX}) and the total resource levels (Q_k).

Allocating Additional Resources When new resources are available, the resource allocation subproblem can be used to determine which facilities the new resources should be allocated to. Thus, in this situation, the levels of one or more resources increase so that $Q_k \geq Q_k^{CUR}$. As well, the level of resources at any facility cannot decrease so that the minimum bound is the current resource allocation, $S_{jk}^{MIN} = S_{jk}^{CUR}$.

Removing Existing Resources Another possible planning scenario would be to examine the impact of removing existing resources from the health care system. The resource allocation subproblem can assist health planners in determining from where to remove these resources so as to have the least negative impact on a system's accessibility. Removing existing resources implies that levels of one or more resources are reduced so that $Q_k \leq Q_k^{CUR}$. As well, since resources are being reduced, the upper bound on the level of resources at each facility is the current allocation so that $S_{jk}^{MAX} = S_{jk}^{CUR}$.

Re-allocating Resources The resource allocation subproblem can also be used to determine whether existing resources can be re-deployed within the health care system so as to increase the aggregate level of accessibility or the equity in the distribution of resources. In this scenario, the total resource levels would remain constant so that $Q_k = Q_k^{CUR}$. Moreover, it is assumed that the resource levels at any given facility must remain within a specific range determined by the decision maker. One possible method of determining this range would be to limit the change in resource levels at any particular facility to be within a particular fraction, F, of the existing allocation so that $S_{jk}^{MIN} = (1 - F)S_{jk}^{CUR}$ and $S_{jk}^{MAX} = (1 + F)S_{jk}^{CUR}$.

Optimal Resource Allocations Finally, the RAS can be used to calculate the optimal deployment of existing resources. The optimal allocations and values of the plan-

ning objectives can then be compared to the existing allocations and current objective function values. For this situation, the total resource levels remain the same and there are no restrictions on the facility resource levels. This implies that $Q_k = Q_k^{CUR}$ and that $S_{jk}^{MIN} = 0$ and $S_{jk}^{MAX} = Q_k$.

It should be emphasized that both the Facility Location Subproblem and the Resource Allocation Subproblem are not the only potential formulations of the AOP. The AOP allows for considerable flexibility in the specification of the objectives circumscribed by a specific health care planning strategy. Further, as demonstrated in the empirical example in Chapter 7, it is quite straightforward to generalize these formulations to reflect a hierarchical model of accessibility.

5.5 Summary

This chapter discussed the use of facility-oriented optimization models to improve the efficiency and equity in the distribution of accessibility among the target population in a health care system. These models take a facility-oriented optimization approach to improving accessibility. This approach proposes that accessibility can be changed by two different strategies: changing the locational configuration of the system and re-allocating resources.

The use of optimization models and facility location models for service development and health care planning has been criticized by some authors. Criticisms of the use of optimization models in health care planning were discussed. This was followed by a discussion of generic optimization objectives, focusing on the concepts of efficiency and equity introduced in Chapter 1. Finally, a generic Accessibility Optimization Problem (AOP) was introduced and partitioned into two tractable subproblems, the Facility Location Subproblem (FLS) and the Resource Allocation Subproblem (RAS). Specialized formulations of both of these subproblems were then provided for four specific scenarios: adding new facilities or resources, removing existing facilities or resources, moving existing facilities or re-allocating existing resources, and computing an optimal system for comparison with the existing system.

The optimization objectives and problems discussed in this chapter are generic, that is, they do not use any particular accessibility measure. The next chapter narrows this generic problem and develops two specific accessibility optimization models using the

minimum-distance accessibility measure and the Joseph and Bantock [1982] accessibility measure, both discussed in detail throughout the thesis.

Chapter 6

Examples of Accessibility Optimization Models

This chapter illustrates the application of the generic AOP using two common measures of geographical accessibility, namely the minimum-distance accessibility measure and the Joseph and Bantock [1982] accessibility measure. For each of these accessibility measures, suitable equity and efficiency objectives are developed which can then be used to generate multiobjective optimization models for the corresponding FLS and, where appropriate, RAS. These optimization problem formulations are illustrated using a small hypothetical example. Finally, specific solution techniques, used in an empirical application in Chapter 7, are introduced for each of these problem formulations.

6.1 Sample Numerical Example

To illustrate the optimization models developed in this chapter, a small problem is used consisting of 25 candidate facility sites with 2 facilities to be located. This example provides a simple concrete application that links back into the concepts and models discussed earlier in the thesis and that demonstrates the multiobjective techniques used in the problem formulations. The trade-off between equity and efficiency and the impact of this trade-off on the model solutions clearly illustrates the issues involved with the problem formulations. The FLS formulations have only 300 feasible solutions so that it is possible to produce a plot of this trade-off between equity and efficiency for

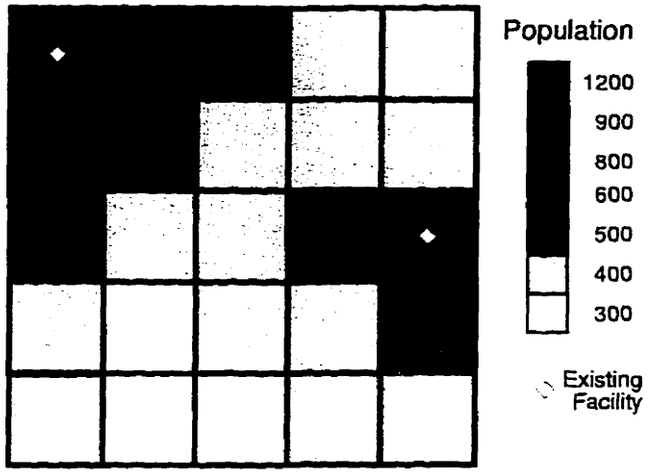


Figure 6.1: System configuration of the example problem. Existing facilities are indicated with a *diamond*.

each solution. Moreover, since there are only two facilities, the equity and efficiency surfaces for the resource allocation subproblem can also be represented easily in a three dimensional graph.

The example problem is shown in Figure 6.1. It consists of 25 sub-areas arranged in a five kilometre by five kilometre grid. Thus, each grid cell is one square kilometre. The population of each grid cell is given in the figure and the total population for this area is 12 100. The problem assumes that the entire population for each grid cell is aggregated to its centre. In order to identify a particular grid cell, the grid cells are numbered in a row-major order. Cells 11, 15, 51, and 55 are at the top left, top right, bottom left, and bottom right corners of the grid respectively. The candidate facility sites are chosen to be the 25 grid cell centres. It is also assumed that currently two equally-sized facilities are located at the centre of grid cells 11 and 35. Finally, all distances between facilities and grid cells are measured using straight-line distances.

It should be emphasized that the results generated by applying the models are for illustrative purposes only and may not generalize in the same way to larger and more realistic problems.

6.2 Minimum Distance Accessibility Measure

The first problem formulation develops a multiobjective optimization model for the minimum distance accessibility measure discussed in Example 2 of Chapter 3. Using this measure, the accessibility of a population sub-group i is defined as

$$A_i = R - \min\left(R, D_{i1}, \ldots, D_{i,N_E}\right) \tag{6.1}$$

where R is the distance beyond which a facility is considered inaccessible and N_E is the number of current facilities. Note that this accessibility measure has a particularly simple form and is only influenced by the facility distances. This leads to two important consequences with respect to the AOP. First, the population only needs to be partitioned spatially into sub-areas or regions. Second, the only way to change the accessibility of the system is by modifying the locations of facilities and, therefore, only a FLS formulation is presented.

6.2.1 Facility Location Subproblem Formulation

Given N_D current facilities and potential facility sites, the goal of the accessibility problem is to determine a locational configuration via the binary vector of decision variables $\mathbf{y}$ so as to maximize one or more planning objectives. This example considers both efficiency and equity objective functions.

For formulation within the optimization model, the accessibility measure defined by equation (6.1) must be modified in order to incorporate the effects of the decision variables, $\mathbf{y}$. Define the function $D(i, \mathbf{y})$ that calculates the minimum distance to a feasible selected facility site for an individual living in a given sub-area i and locational configuration $\mathbf{y}$. A feasible selected facility site is a candidate facility site j that is within range of sub-area i, $D_{ij} < R$, and is currently selected, $y_j = 1$. If there is no feasible selected facility site, then this function returns the maximum range. Thus,

$$D(i, \mathbf{y}) = \begin{cases} \min\limits_{j, y_j = 1} D_{ij} & \text{if } \min\limits_{j, y_j = 1} D_{ij} < R \\ R & \text{otherwise,} \end{cases} \tag{6.2}$$

defines the distance to the nearest selected facility site. Using this notation, the accessi-

bility of a sub-area for a configuration of facilities is

$$A_i = R - D(i, \mathbf{y}) \tag{6.3}$$

and is identical to (6.1) if $y_j = 1$ for all j.

The first planning objective for this problem is the efficiency objective. This objective measures the total aggregate accessibility of the entire target population to the system, as defined by equation (5.1). This corresponds to the sum of the accessibility values of each sub-area weighted by its target population. Mathematically, the efficiency objective function can be expressed as

$$Z_F = \sum_i P_i A_i = \sum_i P_i R - \sum_i P_i D(i, \mathbf{y}). \tag{6.4}$$

The second planning objective relates to the equity of the distribution of accessibility in the target population. For a minimum distance measure, the simplest and most widely applied equity objective is minimizing the maximum distance any individual has to travel to the nearest facility [Mulligan, 1991]. This corresponds to maximizing the accessibility of the least accessible (*i.e., furthest from a facility*) sub-area. Using (6.3), this objective function may be specified as

$$Z_M = \min_i A_i = R - \max_i D(i, \mathbf{y}). \tag{6.5}$$

The goal of this multiobjective optimization problem is to determine the configuration of the candidate facility sites, $\mathbf{y}$, so as to maximize both Z_F and Z_M. As noted in Chapter 2, there are several different techniques for generating non-dominated (efficient) solutions for multiobjective optimization problems. One such technique is the constraint method, where one objective is optimized and the other is constrained to be above a minimum allowable level. The constraint method is an appropriate technique for this problem because the equity objective, Z_M, has an obvious and intuitive interpretation. Namely, $R - Z_M$ can be interpreted as the maximum allowable distance, R', that an individual must travel to reach the nearest facility. Thus, using the constraint method, the optimization problem consists of determining a locational configuration of facilities so as to maximize the aggregate accessibility (or, equivalently, minimize the total distance travelled), given that no individual may travel farther than a specified distance, R'. Furthermore, using this formulation, it is possible for a decision maker

to see the trade-off between minimizing the total distance travelled (efficiency) and the maximum allowable distance from a facility (equity).

Mathematically, this is equivalent to maximizing the efficiency objective, Z_F, and constraining the equity objective so that

$$Z_M \geq \varepsilon_M \tag{6.6}$$

where $\varepsilon_M \geq 0$ is the minimum acceptable accessibility. Note that this is also equivalent to ensuring that the accessibility of every sub-area is greater than the minimum level,

$$A_i = R - D(i, \mathbf{y}) \geq \varepsilon_M \quad \text{for all } i \tag{6.7a}$$

or, equivalently,

$$D(i, \mathbf{y}) \leq R' \quad \text{for all } i \tag{6.7b}$$

where $R' = R - \varepsilon_M$ is the maximum allowable distance to the nearest facility.

After dropping the constant term, $\sum_i P_i R$, from the efficiency objective (6.4) and multiplying by -1, we arrive at the following formulation of the FLS.

$$\underset{\mathbf{y}}{\text{Minimize}} \ \sum_i P_i D(i, \mathbf{y}) \tag{6.7}$$

subject to

$$D(i, \mathbf{y}) \leq R' \tag{6.8}$$

$$\sum_{j=1}^{N_D} y_j = N \tag{6.9}$$

$$y_j \geq Y_j \qquad j = 1, \ldots, N_D \tag{6.10}$$

$$y_j \in \{0, 1\} \qquad j = 1, \ldots, N_D. \tag{6.11}$$

This problem is equivalent to the distance-constrained p-median problem [Choi and Chaudry, 1993].

6.2.2 Numerical Example

The distance-constrained p-median problem is a well-known facility location model, hence the results of applying this model to the example problem are discussed very briefly. Densham and Rushton [1992] provide a detailed example of the application

of this model. Nevertheless, it is useful to apply the model to the sample problem to illustrate the trade-off between equity and efficiency.

From the formulation of the model, the trade-off between equity and efficiency is controlled by the value of the parameter R', the maximum allowable distance to the nearest facility. Furthermore, it is often convenient to compute indicators from the objectives. These indicators give planners information with which to compare the different optimal solutions. For this model formulation, two useful indicators are: the average distance to the nearest facility, $\sum_i P_i D(i,\mathbf{y})/P_T$, and the maximum distance from any sub-area to the nearest facility, $\max_i D(i,\mathbf{y})$.

	Range of R'	Solution	Total Distance	Average Distance	Maximum Distance
Current System		{11,35}	19 630.91	1.6224	4.0000
Optimal System	$R' \geq 3$	{21,34}	16 798.27	1.3883	3.0000
Optimal System	$2.2361 \leq R' < 3$	{31,34}	17 914.33	1.4805	2.2361

Table 6.1: Results of applying the p-median model to the numerical example.

Table 6.1 presents the numerical results of applying the distance constrained p-median model to the sample problem described in Section 6.1. Since there are only 300[1] different feasible solutions, it is possible to evaluate both the total distance and the maximum distance for each solution. These values are plotted in Figure 6.2.

With the existing facilities, located at sites 11 and 35, the average distance to a facility is 1.62 km and the maximum distance from any sub-area to a facility is 4 km. The optimal solution to the p-median model with no maximum distance constraint placed facilities at locations 21 and 34. For this configuration, the average distance was 1.39 km and the maximum distance was 3 km. Thus, the average distance was reduced by approximately 15% and the maximum distance by 25% when compared to the original facility locations. Since the maximum distance of the optimal solution with no maximum distance constraint is 3 km, $\{21,34\}$ is the optimal solution for $R' \geq 3$. By setting R' slightly less than 3, say 2.99 km, the corresponding optimal solution is $\{31,34\}$ with an average distance of 1.48 km and a maximum distance of 2.24 km. Compared to the existing system, this represents a reduction of 44% in the maximum distance and of 9%

[1]The number of ways of choosing 2 sites out of 25.

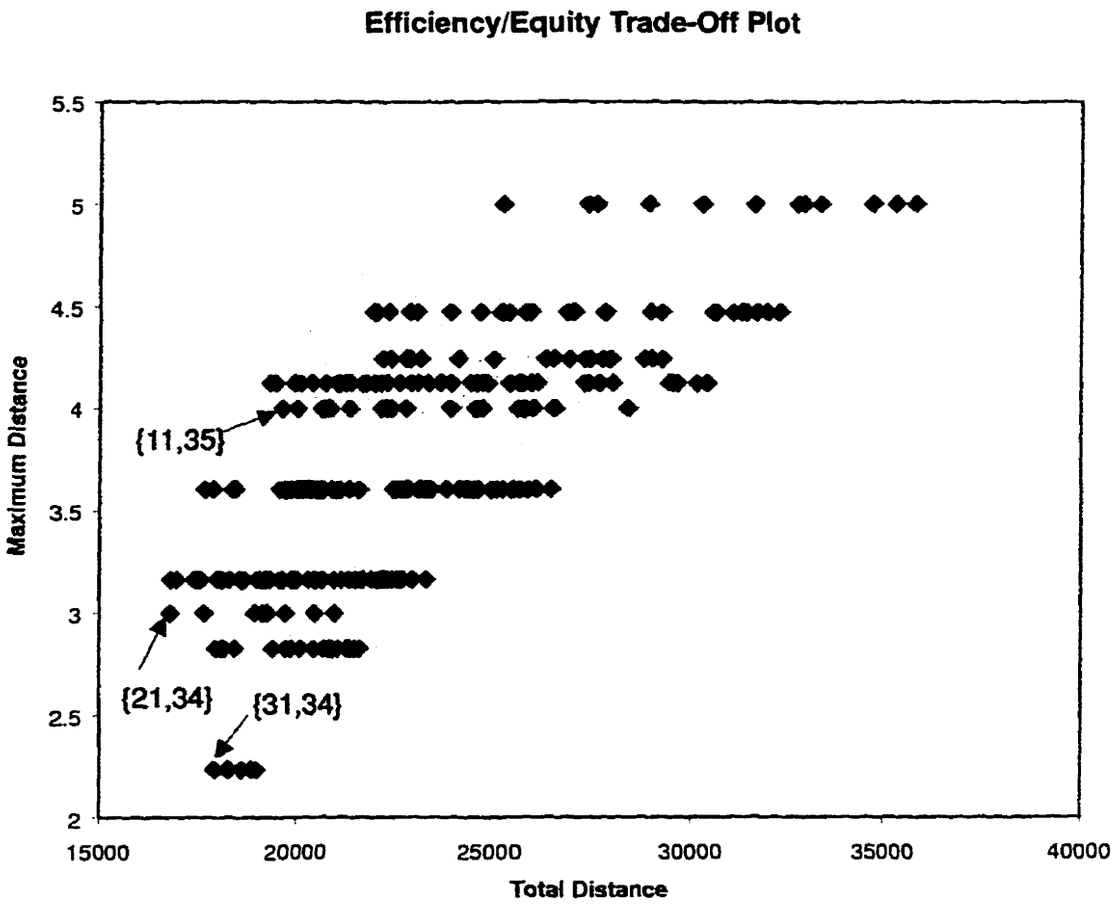


Figure 6.2: Plot of efficiency and equity objective values for every possible solution to the example problem. The existing facility configuration is denoted by $\{11, 35\}$ while the two non-dominated solutions are $\{21, 34\}$ and $\{31, 34\}$.

in the average distance. These two solutions are the only non-dominated solutions for the problem, as there are no optimal solutions for $R' < 2.2361$.

Through inspection of Figure 6.2, it is possible to verify that these are the only two non-dominated solutions as these are the only two points that have no solutions that are both to the left of or below them. Furthermore, since this problem has only 300 possible solutions, it is possible to find the optimal solution through enumeration. For real-world problems where there can be a large number of possible solutions, this is obviously not a feasible approach. The next section discusses methods of solving the distance-constrained p-median model.

6.2.3 Solution Techniques

The distance-constrained p-median problem is also a well-known facility location problem [Khumawala, 1973; Moon and Chaudhry, 1984; Rahman and Smith, 1991; Choi and Chaudry, 1993]. Several different approaches have been suggested for solving this problem. For example, Choi and Chaudhry [1993] used a Lagrangian relaxation and a subgradient method in a branch and bound procedure. Their procedure was tested on several p-median problems with 30 and 150 demand locations. Although Choi and Chaudry reported good results, their method required up to 56 minutes of computer time[2] for these relatively small problems. Therefore, the application of this method for larger problems is questionable in terms of processing efficiency.

Densham and Rushton [1992] use the Interchange (or Teitz and Bart [1968]) heuristic for solving larger distance-constrained p-median problems in a microcomputer-based environment. They used allocation tables, candidate strings, and demand strings to exploit the spatial structure of the problem so as to minimize computation, data volume, and data access times. They also noted that when fixed facilities exist, it is possible to further cull the candidate and demand strings so as to reduce storage space and computation time. Detailed descriptions of these strategies which were tested on several different data sets were provided, and for a large 2844 demand location problem with very tight distance constraints, they reported solution times of under three and a half hours on a slow[3] microcomputer and 13 minutes on a Sun workstation. Thus, the strategies

[2]On a 486 PC running at 33 MHz. Most instances, however, required less time.

[3]A 386 PC computer with a clock speed of 20 MHz.

suggested by Densham and Rushton are reasonable for optimizing minimum-distance accessibility.

6.3 Joseph and Bantock Accessibility Measure

In this second formulation, accessibility is measured using the Joseph and Bantock [1982] (J&B) accessibility measure discussed in Section 2.1.2 and in Example 3 of Chapter 3. As opposed to the minimum distance accessibility measure, the J&B model incorporates the effects that attractiveness and congestion of a facility have on accessibility, in addition to considering the effect of distance. Using this measure, the accessibility of a population sub-group i is defined as

$$A_i = \sum_j (S_j/C_j) \exp(-\beta D_{ij}) \tag{6.12}$$

where S_j is the size of facility j and

$$C_j = \sum_i P_i \exp(-\beta D_{ij})$$

is the congestion of facility j. By defining the constant

$$\gamma_{ij} = \exp(-\beta D_{ij}) / C_j \tag{6.13}$$

accessibility can be expressed as

$$A_i = \sum_j \gamma_{ij} S_j. \tag{6.14}$$

In the following formulation of the AOP the location configuration of facilities and their allocation of resources are used as decision variables.

6.3.1 Problem Formulation

Assume that there is a total of N_D current facilities and potential facility sites. The goal of the optimization problem is to determine both the vector of selected facility locations, $\mathbf{y}$, and facility sizes, $\mathbf{s}$. With these decision variables, the accessibility of a

particular population sub-group i can be calculated as

$$A_i = \sum_{j=1}^{N_D} \gamma_{ij} s_j y_j. \tag{6.15}$$

As with the minimum distance accessibility measure discussed previously, efficiency and equity objective functions are developed for this problem.

This accessibility measure is based on the random utility model discussed in Chapter 4. Hence, an appropriate efficiency objective is to maximize the total satisfaction of the target population. This objective can be expressed as follows.

$$\begin{aligned} Z_S &= \textstyle\sum_i P_i \ln A_i \\ &= \textstyle\sum_i P_i \ln \sum_j \gamma_{ij} s_j y_j. \end{aligned} \tag{6.16}$$

The second objective is the equity objective. One possible measure of the equity is the variance in accessibility of the population, namely,

$$Z_V = \textstyle\sum_i P_i \left(A_i - \overline{A}\right)^2$$

where, for the J&B measure, the average accessibility is

$$\begin{aligned} \overline{A} &= \textstyle\sum_i P_i A_i / \sum_i P_i \\ &= \dfrac{\sum_i P_i \left(\sum_j s_j y_j / C_j\right) \exp\left(-\beta D_{ij}\right)}{\sum_i P_i} \\ &= \dfrac{\sum_j \left(s_j y_j / C_j\right) \sum_i P_i \exp\left(-\beta D_{ij}\right)}{\sum_i P_i}. \end{aligned} \tag{6.17}$$

But $C_j = \sum_i P_i \exp\left(-\beta D_{ij}\right)$, therefore

$$\overline{A} = \textstyle\sum_j s_j y_j / \sum_i P_i = Q / P_T. \tag{6.18}$$

The average accessibility is the total resource level divided by the total target population[4]. Therefore, the equity objective consists of minimizing the variance in accessibility,

[4]This result also demonstrates why the standard efficiency measure cannot be used, as the total aggregate accessibility is constant for any given level of resources and is not affected by the either the location of facilities or the allocation of resources.

namely

$$Z_V = \sum_i P_i \left(\sum_j \gamma_{ij} s_j y_j - Q/P_T \right)^2. \tag{6.19}$$

If, for convenience, we replace the maximization of Z_S with the minimization of $Z_F = -Z_S$, this optimization model involves minimizing both Z_F and Z_V. In order to generate efficient or non-dominated solutions for this problem, this multiobjective problem needs to be transformed into a single objective problem. In comparison with the minimum-distance accessibility optimization problem, neither of these objectives has an obvious interpretation for a planning authority. Furthermore, the equity and efficiency objectives are not necessarily of the same order of magnitude.

However, it is possible to standardize the objective functions since both objectives have known lower bounds. For the equity objective, this lower bound, Z_V^{MIN}, occurs when every $A_i = \overline{A}$ and is obviously zero, corresponding to a completely equitable system with no variance in accessibility. The lower bound for Z_F can also be found by noting that equation (6.18) implies that $\sum_i P_i A_i = Q$ and by finding A_i to minimize Z_F subject to this constraint. The optimal solution[5] is $A_i = \overline{A}$ and, consequently, $Z_F^{MIN} = -P_T \ln \overline{A}$. It is interesting to note that both lower bounds occur with $A_i = \overline{A}$, namely that accessibility is completely evenly distributed within the target population.

As discussed by Malczewski and Ogryczak [1995], one method of generating non-dominated solutions for a multiobjective problem of this type is to use the weighting method. Moreover, it is possible to calculate standardized weights based on the efficiency and equity of the current system, Z_F^{CUR} and Z_V^{CUR}, and their respective lower bounds so that

$$\omega_F = \frac{\omega}{Z_F^{CUR} - Z_F^{MIN}} \quad \text{and} \quad \omega_V = \frac{1-\omega}{Z_V^{CUR} - Z_V^{MIN}} \tag{6.20}$$

where ω, $0 \leq \omega \leq 1$, represents the relative importance placed on the efficiency objective and, consequently, $1-\omega$ is the weighting of the equity objective. Finally, ω_F and ω_V are used to calculate a single objective as the weighted sum of the efficiency and equity

[5]Form the Lagrangian $L = -\sum P_i \ln A_i + \lambda(Q - \sum_i P_i A_i)$. Set $\frac{\partial L}{\partial A_i} = -\frac{P_i}{A_i} - \lambda P_i = 0$ and $\frac{\partial L}{\partial \lambda} = Q - \sum P_i A_i = 0$. Solving this set of equations gives $A_i = Q/\sum_i P_i = \overline{A}$.

objectives. Thus, the objective function for this problem is

$$Z = \omega_F Z_F + \omega_V Z_V.$$

Note that this formulation is convenient for a decision maker as it requires the specification of a single parameter, ω, that represents the trade-off between efficiency and equity.

With this approach, the overall optimization problem can be expressed as follows:

$$\underset{\mathbf{s},\mathbf{y}}{\text{Minimize}} \quad Z = \omega_S Z_F + \omega_V Z_V \tag{6.21}$$

subject to

$$y_j S_j^{MAX} - s_j \geq 0 \qquad j = 1, \ldots, N_D \tag{6.22}$$

$$s_j - y_j S_j^{MIN} \geq 0 \qquad j = 1, \ldots, N_D \tag{6.23}$$

$$\sum_{j=1}^{N_D} s_j \leq Q \tag{6.24}$$

$$y_j \geq Y_j \qquad j = 1, \ldots, N_D \tag{6.25}$$

$$y_j \in \{0,1\} \qquad j = 1, \ldots, N_D \tag{6.26}$$

where S_j^{MIN} and S_j^{MAX} are respectively the minimum and maximum allowable resource levels at candidate facility site j, if a facility is located at that site.

As discussed in the previous chapter, it is possible to partition this optimization problem into the FLS and the RAS by assuming that respectively either $\mathbf{s}$ or $\mathbf{y}$ is fixed. The two subproblems which result from these assumptions are now discussed.

6.3.2 The Facility Location Subproblem

The Facility Location Subproblem (FLS) assumes that the resource level of a facility located at a candidate facility site, if a facility is located at that site, is fixed, *i.e.*, $s_j = S_j = S_j^{MIN} = S_j^{MAX}$. Thus, the optimization problem consists of determining the optimal facility locations, $\mathbf{y}$. With this assumption, the FLS can be formulated as the following

optimization problem.

$$\underset{\mathbf{y}}{\text{Minimize}}\ Z = \omega_S Z_F + \omega_V Z_V \tag{6.27}$$

subject to

$$\sum_{j=1}^{N_D} S_j y_j \leq Q \tag{6.28}$$

$$y_j \geq Y_j \qquad j = 1, \ldots, N_D \tag{6.29}$$

$$y_j \in \{0,1\} \qquad j = 1, \ldots, N_D. \tag{6.30}$$

Numerical Example

This section illustrates the FLS, discussed above, applied to the sample problem defined in Section 6.1. The goal is to select the optimal locations for two equally sized facilities. The total facility size is set to 12 100 so that the average accessibility, $\overline{A}$, is unity . For this problem formulation, the trade-off between equity and efficiency is controlled by the value of the parameter ω which ranges from 0 (pure equity) to 1 (pure efficiency). As a result, the minimum possible value of efficiency, Z_F^{MIN}, is zero, since $Q/P_T = \overline{A} = 1$, as is the lower bound for equity, Z_V^{MIN}. The current values of the two objectives for this system are $Z_F^{CUR} = 793.4$ and $Z_V^{CUR} = 1497.5$. These values were used to standardize the weights using equation (6.20).

As with the p-median formulation, it is possible to define indicators to assist planners and decision makers in assessing the different non-dominated solutions that are generated when the value of ω is varied. One indicator of efficiency is the average satisfaction, Z_{AS}, which is defined as the total aggregate satisfaction divided by the population, $Z_{AS} = Z_S/P_T$, while a useful indicator of equity is the coefficient of variation of accessibility in the target population, Z_{CV}. The maximum possible value for the efficiency is $Z_{CV} = \ln(Q/P_T)$ which, in this case, equals 0. The coefficient of variation is defined as the standard deviation of accessibility divided by the mean and this indicator has the advantage that it can be used to compare equity between different systems. Thus, the coefficient of variation of accessibility can be calculated as $\frac{\sqrt{Z_V/P_T}}{Q/P_T}$. Note that, in this case, Q/P_T equals one so that the coefficient of variation is identical to the standard deviation.

In order to estimate the set of non-dominated solutions of this multiobjective problem, the value of ω was varied from zero to one in increments of 0.05. Three different optimal solutions resulted from these twenty-one values of ω. These solutions are tab-

Weight (ω) Min	Max	Solution	Efficiency (Z_F)	Equity (Z_V)	Avg. Satis.	Coeff. of Variation	Improvement Eff.	Eq.
Current		{11,35}	793.4	1497.5	-0.0656	0.3518	–	–
0.75	1.00	{13,53}	431.10	970.78	-0.0380	0.2751	45.66%	35.17%
0.20	0.70	{22,44}	441.17	924.45	-0.0365	0.2764	44.40%	38.27%
0.00	0.15	{12,44}	460.20	915.87	-0.0356	0.2832	42.00%	38.84%

Table 6.2: Solutions of the facility location subproblem for the numerical example. The percentage improvement values in this table refer to the improvement to the lower bound of each objective.

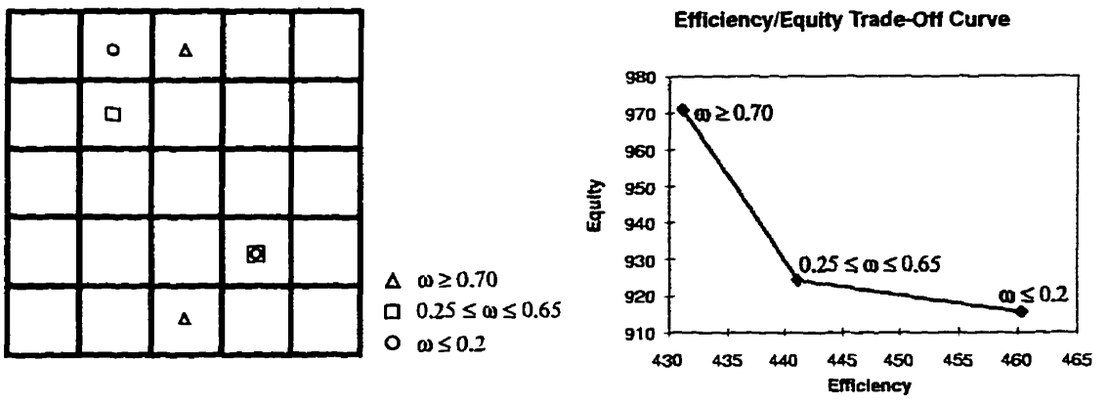
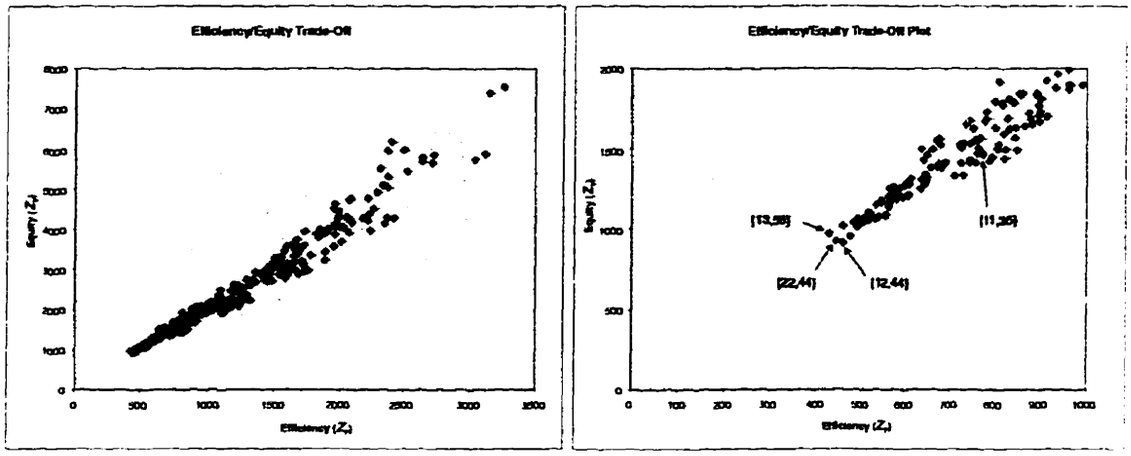


Figure 6.3: Optimal solutions and efficiency-equity trade-off curve for the FLS.

ulated in the fourth, fifth, and sixth rows of Table 6.2 and illustrated in Figure 6.3. The optimal solution was identical to the pure efficiency solution when $\omega \geq 0.75$. When $0.2 \leq \omega \leq 0.7$, the optimal location configuration consisted of facilities located in grid cells 22 and 44. This solution was a trade-off between the efficiency and equity solutions with slightly higher than optimal values of both objectives. Finally, the solution was the same as the pure equity solution when $\omega \leq 0.15$.

For this small problem, it is possible to evaluate the efficiency and equity objectives for each of the 300 feasible solutions; these values are shown in Figure 6.4. One interesting trend visible from this plot is that the two objectives were roughly in agreement, *i.e.*, a solution that had a high value of one objective tended to have a high value of the



(a) All Solutions *(b)* Enlargement of Best Solutions

Figure 6.4: Plot of efficiency and equity objective values for *(a)* all feasible solutions and *(b)* for $Z_F \leq 1000$ and $Z_V \leq 2000$. The existing facility configuration is $\{11, 35\}$ while the three non-dominated solutions are $\{13, 53\}$, $\{22, 44\}$, and $\{12, 44\}$.

other objective. Furthermore, the non-dominated solutions were in a relatively narrow range compared to the objective values of all feasible solutions. This indicates for this small problem that the efficiency and equity objectives are not really conflicting and that a solution that was optimal for one objective was also a good solution for the other.

All of the optimal solutions led to large improvements in both the efficiency and equity objective functions from the existing systems. The corresponding improvement in the average satisfaction was between 42% and 46%, while the reduction in the standard deviation of the accessibility among the target population was between 35% and 40%. Thus, for this problem, the application of the optimization model lead to improvements in both efficiency and equity.

Solution Techniques

As this optimization problem is not a standard facility location model, the standard solution techniques cannot be directly applied. Nevertheless, it is possible to adapt existing heuristic strategies developed for the p-median problem. As mentioned previously, the Interchange (or Teitz and Bart [1968]) heuristic has been used successfully

in solving relatively large scale p-median problems in a microcomputer-based environment [Densham and Rushton, 1992]. Moreover, Birkin *et al.* [1995] report good results using a variant of this heuristic applied to a large scale nonlinear facility location model that involves locating retail facilities so as to maximize market share. Given this, the proposed solution method for this problem is based on the Interchange heuristic.

Recall that the Interchange procedure operates on an existing pattern of selected facilities. At each iteration, a single facility is moved to a vacant site as long as this causes a decrease in the objective function. The procedure terminates when there are no possible moves left that cause a reduction in the objective function. For this FLS formulation, the change in accessibility of population group i caused by moving a facility from site ℓ to a vacant candidate facility site k is

$$\Delta A_i^{\ell k} = \gamma_{ik} S_k - \gamma_{i\ell} S_\ell. \tag{6.31}$$

Therefore, the change in the objective function caused by this interchange can be expressed as

$$\Delta Z_{\ell k} = -\omega_S \sum_i P_i \ln\left(\frac{A_i + \Delta A_i^{\ell k}}{A_i}\right) + \omega_V \sum_i P_i \Delta A_i^{\ell k}\left(2A_i - 2\overline{A} + \Delta A_i^{\ell k}\right), \tag{6.32}$$

where A_i is the unmodified accessibility defined in equation (6.14) and $\overline{A}$ is computed with (6.18). Note that calculating the change in the objective function values does not require the recomputation of the accessibility values. Furthermore, after the interchange, the new accessibility values are equal to $A_i + \Delta A_i^{\ell k}$. This suggests the following heuristic procedure.

1. Given a feasible initial solution $\mathbf{y}$, compute A_i, the initial accessibility values.
2. Set k to the first vacant candidate facility site, where $y_k = 0$, and initialize the number of interchanges for this iteration, N_I, to zero.
3. For each selected facility site, ℓ, that is not a required location[6] and is a feasible interchange with k (*i.e.*, $y_\ell = 1$, $Y_\ell = 0$ and $\sum_j y_j S_j + S_k - S_\ell \leq Q$), calculate $\Delta A_i^{\ell k}$ and $\Delta Z_{\ell k}$ according to equations (6.31) and (6.32) respectively.

[6]Recall that a required location has a 1 in the required locations vector. This indicates that any feasible solution to the FLS must have a facility located at that site.

4. If $\Delta Z_{\ell' k} < 0$ where $\ell' = \arg\min_{\ell} \Delta Z_{\ell k}$, then set $y_k \leftarrow 1$, $y_{\ell'} \leftarrow 0$, increment N_I, and update $A_i \leftarrow A_i + \Delta A_i^{\ell' k}$.

5. If $k < N_D$ then increment k to the next vacant candidate facility site and go to step 3.

6. If $N_I = 0$ (*i.e.*, there were no interchanges during this iteration) then terminate the heuristic, otherwise go to step 2.

The Add and Drop heuristics can also be easily derived for this FLS formulation with the appropriate substitutions into equations (6.31) and (6.32). This allows for a similar strategy to what is often used for the p-median problem, namely applying the heuristics in combination with the Interchange heuristic. Thus, the heuristic methods provide a flexible framework for generating "solutions" to the FLS formulation. However, the FLS assumes that the allocation of resources to a particular site is fixed. The question of allocating resources among existing facilities, examined by the RAS, is discussed in the next section.

6.3.3 The Resource Allocation Subproblem

The resource allocation subproblem assumes that no facilities are opened or closed (*i.e.*, y is fixed) and determines the optimal allocation of resources among the existing facilities. With this assumption, the RAS can be formulated as follows.

$$\underset{\mathbf{s}}{\text{Minimize}} \; Z = \omega_S Z_F + \omega_V Z_V \tag{6.33}$$

subject to

$$\sum_{j=1}^{N_D} s_j \leq Q \tag{6.34}$$

$$S_j^{MIN} \leq s_j \leq S_j^{MAX} \quad j = 1, \ldots, N_D. \tag{6.35}$$

This is a nonlinear programming problem with a convex feasible region defined by one linear constraint and by simple bounds on every variable. For the RAS, it is important to note that there are N_D continuous decision variables and no binary variables. Thus, this problem lacks the combinatorial aspects of the AOP and the FLS and can be solved with standard nonlinear optimization methods (see, for example, [Gill *et al.*, 1981; Luenberger, 1984] for a thorough discussion of these methods).

Numerical Example

The RAS, as applied to the sample problem outlined in Section 6.1, consists of determining the optimal sizes of the existing facilities located at grid cells 11 and 35. For this problem, the decision variables representing the facility sizes are denoted by s_{11} and s_{35}. As before, the maximum total facility size, Q, is 12 100 and, consequently, the average accessibility, $\overline{A}$, is 1. Figure 6.5 illustrates this problem with a three-dimensional surface plot and contour diagrams for both the efficiency and equity objectives. Note that the original efficiency objective, $Z_S = -Z_F$, is plotted as the efficiency surface to enhance the legibility of its surface plot. The axes of the plane, labelled s_{11} and s_{35}, represent the sizes of the two facilities. The vertical axis represents the value of the corresponding objective in arbitrary units. Finally, the diagonal line on the two contour diagrams represents $s_{11} + s_{35} = 12\,100$ and, with the non-negativity constraints, defines the feasible region for this problem as the triangle below and to the left of this line.

The minima of the efficiency objective, Z_F^*, and equity objective, Z_V^*, are indicated on the two contour diagrams, as are the corresponding optimal facility sizes, s_{11}^* and s_{35}^*[7]. The objective weights were calculated as in the previous example. In order to examine the efficiency-equity trade-off, ω was varied from zero (pure equity) to one (pure efficiency) in increments of 0.1. These values are also shown in Table 6.3.

One interesting point to note in Table 6.3 is that the total size for the optimal solution for $\omega \leq 0.1$ is less than the maximum allowable size. In fact, for the pure equity case, the total facility size is only 10 840 compared to the limit of 12 100. Thus, the optimal solution for the pure equity case allocates fewer than the total available resources. However, this objective does not measure the actual variance in the accessibility because the average accessibility, $\overline{A}$, was fixed at one. The true minimum of the equity objective would occur when $\mathbf{s} = 0$, *i.e.*, when no resources are allocated to any facility. This is a perfectly equitable solution as the system is equally accessible (completely inaccessible) to every member of the target population. Nevertheless, this is neither a desirable nor a reasonable solution to the problem.

The logical conclusion from these results is that this problem is misspecified. One way to correct this misspecification is to modify constraint (6.34) to be an equality con-

[7]For the situation where constraint (6.34) is an inequality constraint. The equality-constrained subproblem is discussed subsequently.

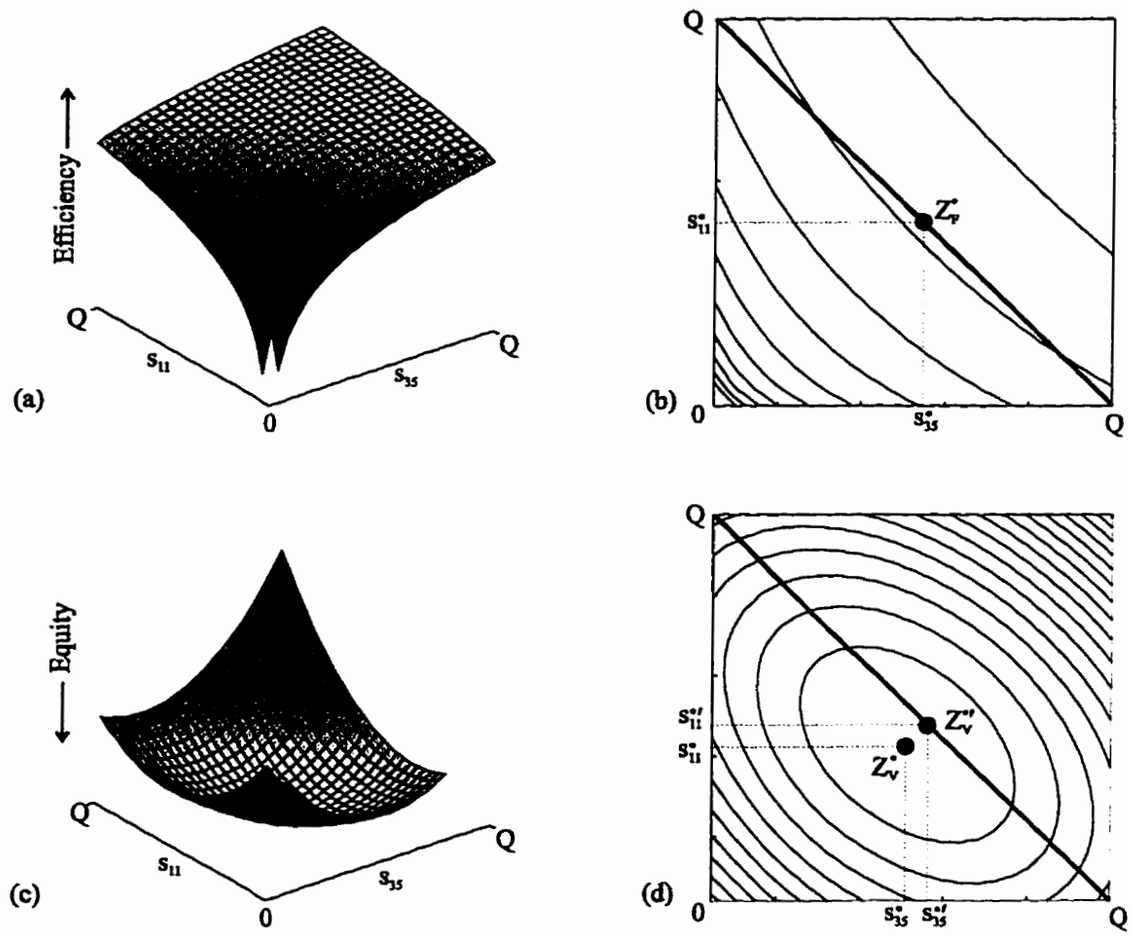


Figure 6.5: Three-dimensional surfaces and contours of the efficiency and equity objectives.

Weight	Efficiency	Equity	Avg.	Coeff. of	Improvement		Facility Sizes		
ω	Z_F	Z_V	Satis.	Variation	Eff.	Eq.	s_{11}	s_{35}	Total
Current	793.4	1497.5	-0.0656	0.3518	–	–	6050	6050	12100
0.0	2070.1	1259.9	-0.1711	0.3227	-167.98%	13.74%	4875	5965	10840
0.1	852.4	1381.4	-0.0704	0.3379	-10.35%	5.43%	5374	6613	11987
0.2	738.9	1406.5	-0.0611	0.3409	4.35%	3.70%	5407	6693	12100
0.3	738.5	1406.8	-0.0610	0.3410	4.41%	3.69%	5387	6713	12100
0.4	738.0	1407.2	-0.0610	0.3410	4.46%	3.66%	5365	6735	12100
0.5	737.6	1407.8	-0.0610	0.3411	4.51%	3.61%	5341	6759	12100
0.6	737.2	1408.7	-0.0609	0.3412	4.56%	3.55%	5314	6786	12100
0.7	736.9	1410.0	-0.0609	0.3414	4.61%	3.47%	5285	6815	12100
0.8	736.6	1411.6	-0.0609	0.3416	4.65%	3.35%	5252	6848	12100
0.9	736.4	1413.9	-0.0609	0.3418	4.67%	3.20%	5215	6885	12100
1.0	736.3	1416.9	-0.0609	0.3422	4.68%	2.99%	5174	6926	12100

Table 6.3: Objective function values and allocations for the resource allocation subproblem with inequality constraint.

straint so that

$$\sum_{j=1}^{N_D} s_j = Q. \tag{6.36}$$

This constraint will then ensure that the total facility size is equal to the maximum allowable level of resources. Thus, the feasible region for the equality-constrained problem is along the diagonal line $s_{11} + s_{35} = 12\,100$, as illustrated in the contour diagrams of Figure 6.5. The optimal pure equity solution for the equality-constrained problem must lie on this line and, consequently, is at the point labelled $Z_V^{*\prime}$. Table 6.4 contains the results of re-evaluating the efficiency-equity trade-off curve for the equality-constrained problem.

In contrast to the inequality-constrained problem, the range of the objective functions is much more restricted. The difference in average satisfaction and the standard deviation of accessibility at the pure equity and pure efficiency solutions is less than 0.5%. Hence, the efficiency and equity solutions are in fairly close agreement. Nevertheless, the efficiency objective was reduced by slightly over 4% and the equity objective by around 3.5% from the initial equally-weighted solution. This indicates that, for this sample problem, it was possible to make at least modest gains in the average satisfac-

Weight	Efficiency	Equity	Avg.	Coeff. of	Improvement		Facility Sizes		
ω	Z_F	Z_V	Satis.	Variation	Eff.	Eq.	s_{11}	s_{35}	Total
0.0	739.7	1406.3	-0.0611	0.3409	4.24%	3.72%	5442	6658	12100
0.1	739.3	1406.4	-0.0611	0.3409	4.30%	3.71%	5425	6675	12100
0.2	738.9	1406.5	-0.0611	0.3409	4.35%	3.70%	5407	6693	12100
0.3	738.5	1406.8	-0.0610	0.3410	4.41%	3.69%	5387	6713	12100
0.4	738.0	1407.2	-0.0610	0.3410	4.46%	3.66%	5365	6735	12100
0.5	737.6	1407.8	-0.0610	0.3411	4.51%	3.61%	5341	6759	12100
0.6	737.2	1408.7	-0.0609	0.3412	4.56%	3.55%	5314	6786	12100
0.7	736.9	1410.0	-0.0609	0.3414	4.61%	3.47%	5285	6815	12100
0.8	736.6	1411.6	-0.0609	0.3416	4.65%	3.35%	5252	6848	12100
0.9	736.4	1413.9	-0.0609	0.3418	4.67%	3.20%	5215	6885	12100
1.0	736.3	1416.9	-0.0609	0.3422	4.68%	2.99%	5174	6926	12100

Table 6.4: Objective function values and allocation for the equality-constrained resource allocation subproblem.

tion and to make reductions in the coefficient of variation in the target population by re-allocating resources. The behaviour of the RAS is not necessarily indicative of its usefulness when applied to real-world systems. Although it is possible to graph and visualize the RAS when there are only 2 decision variables, this becomes much more difficult when dealing with a large number of facilities. The non-linear programming techniques, discussed below, are used to solve larger-sized resource allocation subproblems typical in real-world applications. The results of their application are presented and discussed in Chapter 7.

Solution Techniques

This resource allocation subproblem has a specific form that allows for its efficient solution. The objective function, Z, is a smooth or twice-continuously differentiable function. For

$$\gamma_i^T = [\gamma_{i1}, \ldots, \gamma_{i,N_D}] \tag{6.37}$$

the gradient vector, $\mathbf{g}$, of Z is equal to

$$\mathbf{g} = \sum_i P_i \left[-\omega_S / A_i + 2\omega_V \left(A_i - \overline{A}\right)\right] \gamma_i \tag{6.38}$$

while the Hessian matrix (the matrix of second partial derivatives) is

$$\mathbf{H} = \sum_i \kappa_i \gamma_i \gamma_i^T \tag{6.39}$$

where $\kappa_i = P_i\left(\omega_S/A_i + 2\omega_V\right)$ is defined for notational convenience. Since P_i and $\omega_S + \omega_V$ are positive, κ_i is also positive. However, note that the gradient vector and the Hessian matrix contain A_i^{-1} terms and therefore become undefined if $A_i = 0$ for any i.

Since the Hessian matrix is positive semi-definite[8], the objective function is convex. Therefore, any local minimum within the feasible region of the problem is also a global minimum of the constrained problem [Luenberger, 1984, p. 181]. Moreover, this property can often be strengthened because if the vectors γ_i span $\mathbb{R}^{N_D}$, then the Hessian matrix is positive definite[9], in which case the optimization problem has a single unique global minimum. Due to these properties, this optimization problem can be considered a well-behaved problem as long as $A_i > 0$. If $A_i = 0$ for any i and $\omega_S > 0$, then the objective function, the gradient vector, and the Hessian matrix are undefined. However, $A_i = 0$ can only occur in two situations. First, if $\gamma_i = 0$, then this sub-group can be removed for the problem as the RAS cannot improve accessibility for sub-groups that are out of range of every facility. The second situation occurs when $\sum_j \gamma_{ij} s_j = 0$ and $\gamma_i \neq 0$. This situation can be remedied by defining the lower bound on the allocation of resources to be positive so that $S_j^{MIN} > 0$. Of course, computational problems may occur if A_i is very small.

A good solution method for this convex problem is a constrained form of Newton's method. Given a feasible starting point, $\mathbf{s}^{(0)}$, set $\ell \leftarrow 0$, and repeat the following steps:

1. Terminate the algorithm if convergence conditions are satisfied at $\mathbf{s}^{(\ell)}$. These convergence conditions test whether $\mathbf{s}^{(\ell)}$ is a constrained minimum of the FLS (see, for example, [Gill *et al.*, 1981, p. 308]).

2. Compute the search direction $\mathbf{p}^{(\ell)}$. The search direction can be found by minimizing a quadratic model based on the Taylor-series expansion of the objective function about the current point, $\mathbf{s}^{(\ell)}$. The search direction is the solution to the

[8]For any vector $\mathbf{x} \in \mathbb{R}^{N_D}$, $\mathbf{x}^T\mathbf{H}\mathbf{x} = \sum_i \kappa_i \left(\gamma_i^T \mathbf{x}\right)^2 \geq 0$, so that the matrix is positive semi-definite by definition. See Gill *et al.* [1984] for a further discussion on the definiteness of a matrix and its implications in optimization.

[9]Since the vectors span $\mathbb{R}^{N_D}$, $\gamma_i^T \mathbf{x} > 0$ at least once for $\mathbf{x} \in \mathbb{R}^{N_D}$, $\mathbf{x} \neq \mathbf{0}$. Therefore $\mathbf{x}^T\mathbf{H}\mathbf{x} > 0$ and thus, by definition, the Hessian is positive definite.

following constrained Quadratic Programming (QP) problem.

$$\underset{\mathbf{p}^{(\ell)}}{\text{Minimize}}\ \mathbf{p}^{(\ell)T}\mathbf{g}\left(\mathbf{s}^{(\ell)}\right) + \frac{1}{2}\mathbf{p}^{(\ell)T}\mathbf{H}\left(\mathbf{s}^{(\ell)}\right)\mathbf{p}^{(\ell)} \tag{6.40}$$

subject to

$$\sum_j p_j^{(\ell)} = 0 \tag{6.41}$$

$$S_j^{MIN} - s_j^{(\ell)} \leq p_j^{(\ell)} \leq S_j^{MAX} - s_j^{(\ell)} \qquad j = 1, \ldots, N_D. \tag{6.42}$$

Since the Hessian matrix is positive semi-definite, this constrained QP problem possesses a minimum which can be computed efficiently [Gill *et al.*, 1981, p. 177].

3. Calculate the step length $0 < \theta^{(\ell)} \leq 1$ so that there is a "sufficient decrease" in the objective function Z. The calculation of $\theta^{(\ell)}$ is discussed further below.

4. Set $\mathbf{s}^{(\ell+1)} \leftarrow \mathbf{s}^{(\ell)} + \theta^{(\ell)}\mathbf{p}^{(\ell)}$, $\ell \leftarrow \ell + 1$, and go back to step 1.

As noted by Gill *et al.* [1981, p. 100], there must be a "sufficient" decrease in the objective function in order to ensure that the model algorithm, outlined above, converges. One way of meeting this condition is to ensure that the step length satisfies the Goldstein-Armijo principle[10]:

$$0 < -\mu_1\theta\mathbf{g}(\mathbf{s})^T\mathbf{p} \leq Z(\mathbf{s}) - Z(\mathbf{s} + \theta\mathbf{p}) \leq -\mu_2\theta\mathbf{g}(\mathbf{s})^T\mathbf{p} \tag{6.43}$$

where $0 < \mu_1 \leq \mu_2 < 1$. Thus, one method of calculating the step length would be to set θ to be the first member of the sequence $1, 0.5, 0.25, \ldots$ which satisfies (6.43). Further details on step length algorithms can be found in Gill *et al.* [1981].

If the objective function consists of only the equity objective, *i.e.*, $\omega_S = 0$, then the Hessian matrix is constant and the quadratic model of this function is exact. Thus, solving the quadratic model, defined by equations (6.40) to (6.42), and using a step length of 1 yields the exact solution to the original problem and the procedure terminates after one iteration. Thus, in comparison to the facility location subproblem, this optimization problem can be solved efficiently, particularly for the equity-only case.

[10]The iteration counter (ℓ) has been omitted for notational convenience.

6.4 Summary

This chapter provided two specific examples of the Accessibility Optimization Problem (AOP) for the minimum distance accessibility measure and the Joseph and Bantock [1982] accessibility measure. The properties of the derived optimization models are explored using a small sample problem.

The first example used one of the simplest accessibility measures, namely the minimum-distance accessibility measure. For this measure, accessibility is based only on the distance to the nearest facility and does not consider the effect of resource levels on accessibility. Consequently, the RAS is not appropriate for this accessibility measure and the AOP is equivalent to the FLS. An efficiency objective based on maximizing the total accessibility and an equity objective consisting of maximizing the minimum accessibility were derived. Using these objectives, the AOP can be shown to be equivalent to the distance-constrained p-median problem.

The second example used the Joseph and Bantock [1982] accessibility measure to develop a more complicated optimization model. This measure is influenced by the distances to facilities and by their size or resource levels as well as by the potential demand on the facilities. Both an efficiency measure, based on maximizing the aggregate satisfaction, and an equity measure, which minimizes the variance in accessibility, were developed for this model. A FLS formulation and a RAS formulation were then applied to the example problem to examine the effects on efficiency and equity of relocating facilities and of modifying facility sizes. A modified Interchange heuristic was proposed for the FLS location subproblem, while a constrained Newton's method algorithm was proposed for the RAS.

The optimization models developed in this chapter were demonstrated on a small sample problem. Although this problem was convenient to illustrate these models, it is important to illustrate their application on a real-world data set. The next chapter applies these accessibility measures to a data set of greater complexity than the example problem to evaluate accessibility to health care services. This application focuses on the accessibility of women in the fertile age cohort to family planning services in the Central Valley of Costa Rica.

Chapter 7

Applying Accessibility Evaluation and Optimization Models

This chapter applies the accessibility measures and optimization models presented in the previous chapter to examine the accessibility of women in the fertile age cohort to family planning services in the Central Valley of Costa Rica. Three types of health care facilities offer family planning, namely, hospitals, clinics, and health centres. To accommodate these three distinct types of health care facilities, the optimization models and solution techniques developed in previous chapter are modified to consider this three-level service hierarchy. Further, the population of the study area is partitioned into three sub-groups based on area of residence: urban, suburban, and rural. With these three sub-groups, it is possible to examine differential accessibility among these groups and, moreover, to examine the impacts that the optimal solutions have on the accessibility.

The objective of this chapter is to test empirically the proposed accessibility optimization models using real world data. Consequently, the main emphasis is on applying the models using different sets of parameters. Two optimization scenarios are applied to the data set for each optimization model. The first scenario determines a fully optimized configuration or resource allocation of the existing system. A second scenario applies the optimization models to determine where additional resources or facilities should be located.

The next section describes the context of this problem and describes the specific study area and the associated geographic data layers. Next, the implications that this

data set has on the optimization models is examined. The subsequent sections apply the models to these data. First, the minimum distance accessibility of the study area is examined and the two optimization scenarios are applied using a distance-constrained p-median model. Next, the Joseph and Bantock [1982] model is considered. Data from the 1992 Costa Rican Reproductive Health Survey are used to calibrate this model. The results of this calibration are then used to evaluate the accessibility of family planning services in the Central Valley. Finally, both the Facility Location Subproblem (FLS) and the Resource Allocation Subproblem (RAS) formulations, discussed in the previous chapter, are applied for both optimization scenarios.

7.1 Accessibility to Family Planning Services in the Central Valley of Costa Rica

Although fertility in Central America is on the decline in general through changes in reproductive behaviour, including increased use of contraception and family planning [Guzman, 1992], socio-economic segregation of the population and the geographic concentration of poor households into service-poor communities is now a widespread phenomenon throughout the region. Rural-urban migration has changed the traditional child-bearing and domestic roles of women through their increased participation in the formal and informal urban labour markets.

However, in countries such as El Salvador and Nicaragua, which are in the early stages of demographic transition, fertility rates are still high and women bear on average more than 4.5 children during their fertile years. In more economically advanced countries, such as Costa Rica, fertility rates and average family sizes are more in line with the developed world. Evidence, assembled by CELADE [1992] and Chakiel and Martinez [1992], indicates that fertility in all countries in the region is much higher in rural than urban areas, as contraceptive use and family planning practices are less prevalent and availability of health care is generally poor.

Current accessibility, both geographic and socio-economic, of maternal health care and family planning (MHC/FP) services is now better overall than ever before in most Central American countries but, as with many other goods and services, health care access and use remains highly variable, especially in rural areas. It is estimated by ECLAC [1992] that 130 million people throughout the whole of Latin America do not

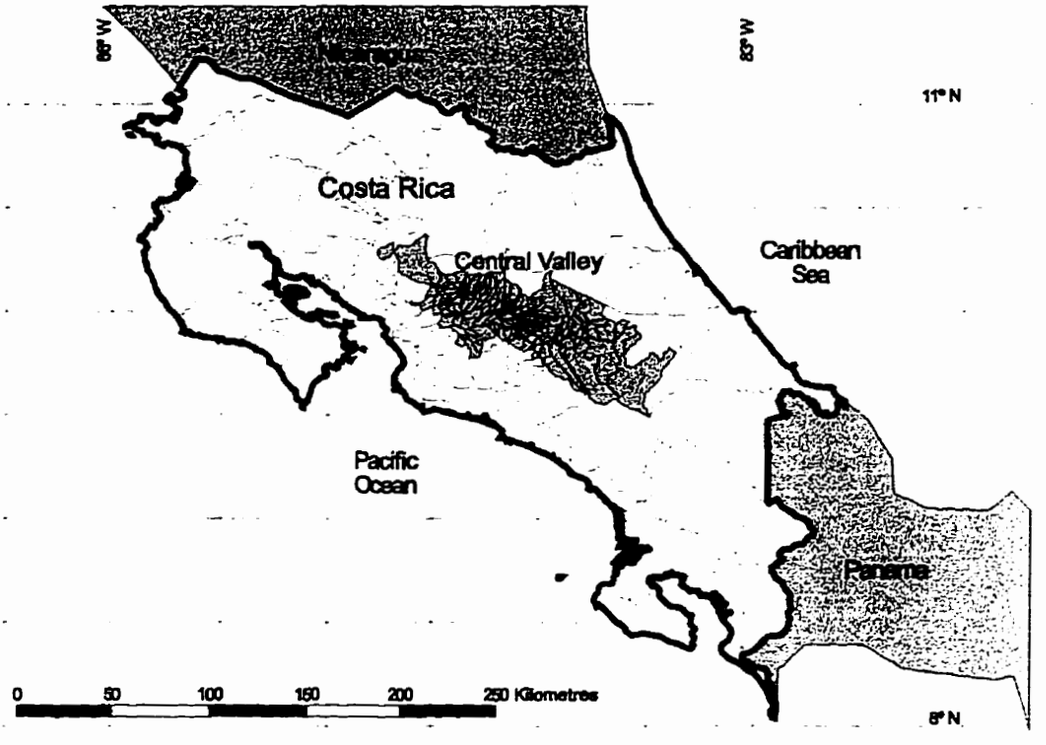


Figure 7.1: Geographic extent of the study area.

have even minimally acceptable access to health services and, of this number, 90 million are women of child-bearing age and children.

The accessibility evaluation and optimization models discussed previously are applied to a data set that pertains to the provision of family planning services in the Central Valley of Costa Rica. Although the boundary of the Central Valley is loosely defined and does not correspond to any official or aggregate administrative district, the base area for this analysis is a spatial data set consisting of the geographic boundaries of 209 *distritos* (districts) in the Central Valley, which, at the time of the 1984 census (shown in Figure 7.1), had a total census population of 1 456 614. The East-West extent of the study region is approximately 150 km while the North-South extent is about 100 km and the total area contained is 4935 square kilometres.

The base year for the analysis was chosen to be 1992, the year of the most recent Reproductive Health Survey [CCSS, 1994]. However, 1984 is the most recent year for

which census data are available. The number of women in the fertile age cohort in 1992 were estimated from the population of women aged 7 to 41 in 1984. These estimates ignore the effects of births, deaths, and migrations. Neglecting births and deaths has a minimal impact on the estimates[1].

A more critical issue in this context is the effect of migration. Migration can have a potentially large impact on the distribution of population. The study area includes the city of San José, not only the capital city of the country, but also the largest city. ECLAC [1993] provides information that the San José metropolitan area grew at an annual rate of 4.21% from 1970 to 1990 compared with a growth rate of 2.77% for the country as a whole. Thus, San José is a likely destination for in-migration from rural areas of the country. Further, a large proportion of the in-migrants to the San José region are likely to be in the 15 to 30 year old portion of the target population. It is highly likely, therefore, that there is a substantial underestimation of the target population in the outer suburbs of San José, where population growth is taking place, due to the higher mobility of a portion of target population and the effect of in-migration. Unfortunately, it was not possible to obtain more recent population estimates as the population figures at the distrito level produced by the Costa Rican Census Department between census years also ignore the effects of migration (for example, see Dirección General de Estadísticas y Censos [1992]). Nevertheless, as mentioned previously, the main objective of this chapter is to illustrate the application of the accessibility evaluation and optimization models and not to perform a comprehensive examination of accessibility in the study area. Therefore, although these estimation errors undoubtedly influence the accessibility evaluation and optimization analyses, this data set more than suffices for illustrating the models.

7.1.1 Population

The first data layer in this data set was the base population layer. The base population for this analysis was taken to be women in the fertile age cohort in 1992, defined by the CCSS [1994] as 15-49 year olds. These women were further partitioned into three mutually exclusive and exhaustive sub-groups by their area of residency, namely urban, suburban, or rural. These three population groups were identified in order to

[1] People born since 1984 are not included in the 1992 fertile cohort and the death rate for this age group is low. The life expectancy for women in Costa Rica in 1985-90 is 77 years [Ross *et al.*, 1992].

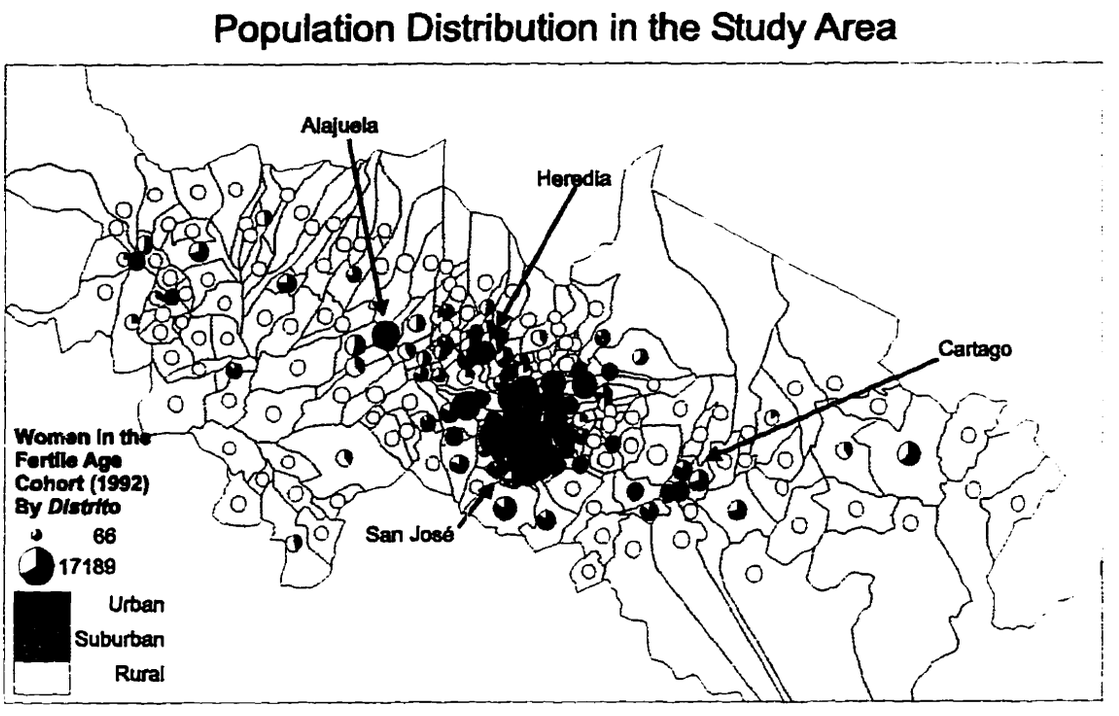


Figure 7.2: Distritos in the Central Valley of Costa Rica with the estimated number of women in the fertile age cohort in 1992 partitioned by area of residency. The spot charts are located at the approximate demographic centroid of each distrito.

examine the existing family planning accessibility levels and assess the impact of the optimization models on differential geographic accessibility.

The population counts of urban, suburban, and rural women aged 7 to 41 (in 1984) in each distrito as well as a distrito identifier were extracted from the Costa Rican census database and, as mentioned previously, were used as estimates of women in the fertile age cohort in 1992. Using these estimates, 283 107 women resided in urban areas, 32 319 in suburban areas, and 151 268 in rural areas for a total of 466 694 women. The number of women in each distrito varied from a minimum of 66 to a maximum of 17 189. These population figures were linked to the equivalent distrito polygon in the distrito layer of a Geographic Information System (GIS) database. The spatial distribution of women in the fertile age cohort within the study area is shown in Figure 7.2. As expected, the main population concentration is around the capital city of San José with smaller concentrations in Alajuela, Heredia, and Cartago.

The definitions of the urban, suburban, and rural population groups were derived from an external variable in the 1984 census relating to zone of residence. In the census, the settlement pattern of each *segmento*[2] was classified as one of urban, suburban, concentrated rural, or dispersed rural. As the locations of the segmentos were not available for this analysis, the total population of each distrito was obtained in the three population groups (with the two rural classes merged into one group). The definitions used by the census authority to generate these classification were not available, however, as shown in Figure 7.2, the main concentration of urban populations are in the major cities while the suburban population is concentrated around the periphery of San José. Again, it should be noted that the population of women in the fertile age groups within these peripheral areas is most likely substantially higher that the estimates due to the effects of in-migration.

Nevertheless, there are always uncertainty and questions of accuracy when dealing with secondary data. For example, the suburban population is relatively small when compared to the two other population groups. It is unclear why the suburban population is so small since it is not known how there population groups were defined by the Costa Rican census authority. This could be related to the actual settlement pattern or could be a result of the definitions used. In lieu of using the census definitions, it could be possible to define these three classes by the population density.

Each distrito, especially at the periphery of the study area, describes a fairly large, irregularly-shaped, and non-homogeneous geographic area that has the potential for a high level of spatial aggregation error as discussed in Chapter 4. The smallest distrito had an area of 0.75 km^2 while largest was 664 km^2. The average area is 23.6 km^2 and the median area is 8.86 km^2. Furthermore, the peripheral distritos not only vary widely in area but are located in relatively sparsely populated mountainous regions. Given this variation in area among the distritos, the level of spatial aggregation can vary widely between distritos. The cumulative effect of this variation is to produce potentially larger errors in measuring the distances between service consumers and the service supply points. Without appropriate adjustment, this error can render the results of an accessibility analysis, using such spatially aggregated data, virtually meaningless.

In order to overcome these difficulties, the approach proposed in Chapter 4 was used and the irregularly shaped polygons were transformed to a raster or regular grid-

[2]The census division underneath *distrito* and containing approximately 200 people.

cell-based representation of the target population. As mentioned previously, there are two advantages to this approach. First, each grid-cell has the same size and shape, which has the tendency to reduce the variation in aggregation error between grid-cells. Second, the size of the grid-cells can be chosen so as to achieve a desired level of accuracy. For this analysis, a 750 metre square grid-cell is used. From equation (4.12), the error bound in estimating distance using a 750 metre grid-cell is 0.287 km for the minimum distance accessibility measure. Similarly from equation (4.24), the error bound in the estimation of the distance deterrence, F_k^A, with a raster of this area is bounded by

$$0.8676 \leq F_k^A \leq 1.1559$$

with a decay parameter of 0.5. Thus, the maximum error in calculating distance deterrence is approximately 15%.

The population points for the distrito polygon to grid disaggregation were obtained by estimating a demographic centroid for each distrito. For the majority of distritos, this was taken to be its centroid. However, in the large distritos at the periphery of the study area, these points were located near the major town or populated area in that distrito. The spot charts in Figure 7.2 are centred on these population points. Since no information of the land use for the grid-cells was available, the population propensity, discussed on page 100 in Section 4.3.2, for each cell was unity as was the α parameter. Finally, the maximum radius of the "window" in the disaggregation procedure, R, was set to 5.5 km, which is approximately equal to the diameter of a circle of area 23.6 km^2, the average area of the 209 distritos. This window approximates the distance between the centroids of two adjacent circular distritos. The distrito polygons were used to adjust the urban, suburban, and rural populations to ensure that the total population of the grid-cells within a particular distrito matched the distrito population. It should be emphasized, however, that although the level of potential spatial aggregation error is reduced, the population counts at the grid-cell level are only estimates. It is likely that the actual population counts vary substantially from these estimates.

This procedure resulted in a subsequent grid-cell data layer consisting of 5028 populated grid cells, illustrated in Figure 7.3. Although errors in grid populations undoubtedly exist in the database, nonetheless, it provides a reasonable approximation to the actual population distribution, in lieu of more spatially disaggregate population data, for the purposes of testing the models.

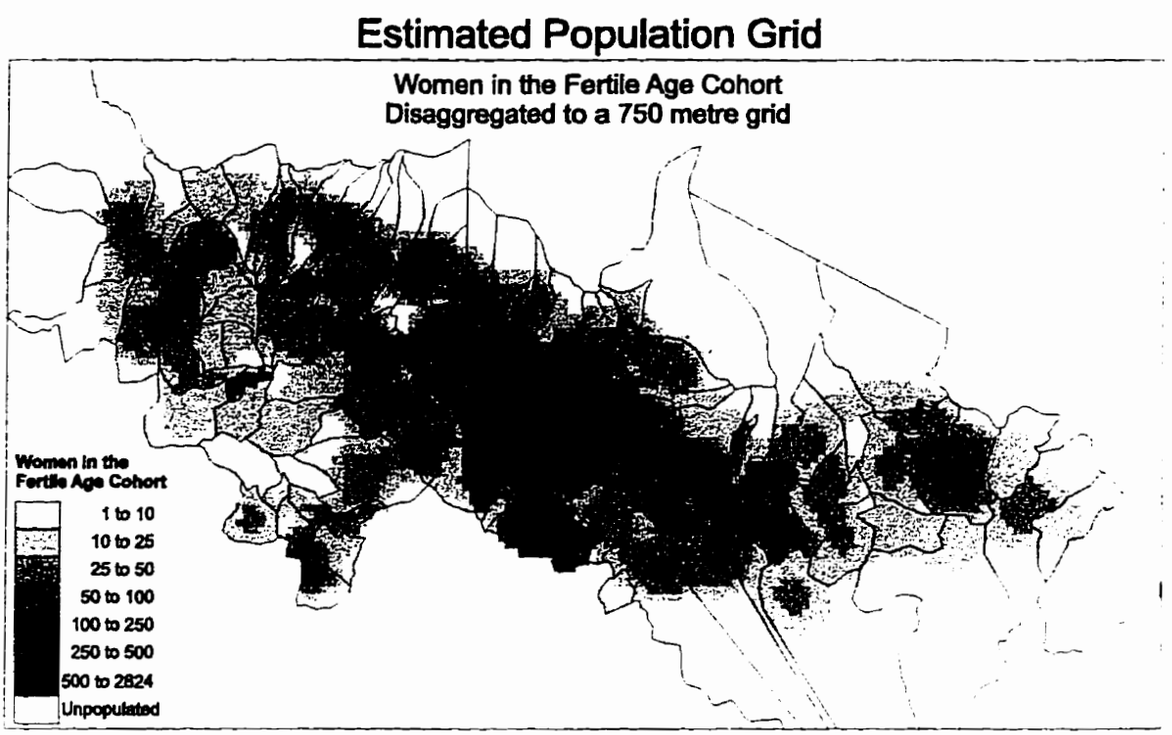


Figure 7.3: Total number of women in the 1992 fertile age cohort disaggregated to a 750 metre grid.

7.1.2 Facilities

The second data layer contains information on the spatial distribution of the service providers. The public sector is the main provider of family planning in Costa Rica through hospitals and clinics provided by the Department of Social Security (CCSS) and through health centres and health posts run by the Ministry of Health. These facilities provide over 75% of modern contraception in Costa Rica [Rosero, 1995].

The service provider data layer consists of 120 delivery points of public family planning services, coded by type of facility, located within the geographic boundary of the study area. This layer contains the geographic locations of 14 hospitals and 58 clinics run by the CCSS, and 48 health centres run by the Costa Rican Ministry of Health. Private sector sources of health service supply, such as pharmacies and private physicians, were not included in these data. Further, health posts, run by the Ministry of Health were also excluded as they provide few family planning services [Rosero, 1995]. The

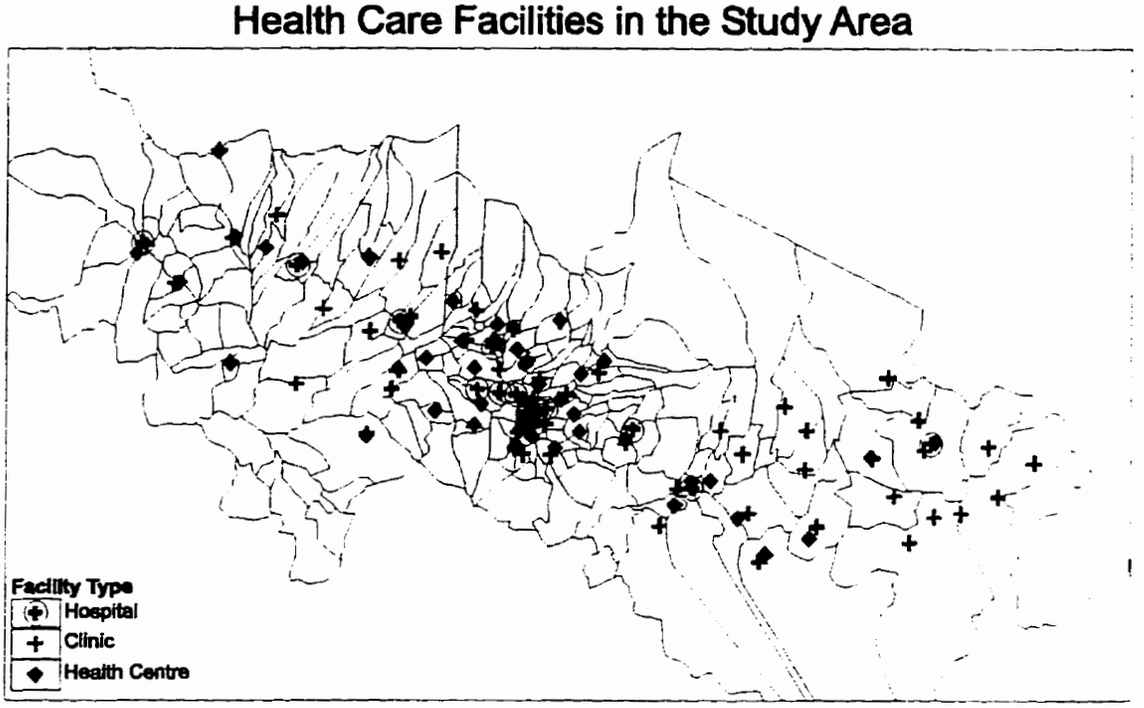


Figure 7.4: Locations of service delivery points in the study area.

locations of the 120 service delivery points are shown in Figure 7.4.

The number of annual hours of family planning service consultations in 1992 were obtained for each of the 120 facilities from Dr. Luis Rosero of the University of Costa Rica, based on unpublished data on outpatient consultations. These figures were used as a surrogate for the resource availability at a facility required for the Joseph and Bantock accessibility measure.

7.1.3 Road Network

The third geographic data layer consists of the road network in the study area. All roads in rural areas and major urban roads within the study area were digitized and coded by road type from 1:50 000 topographical map sheets obtained from the National Geographic Institute of Costa Rica. In order to connect the network, roads that ran between sections of the study area but outside its boundary were also digitized and coded. This road network shown in Figure 7.5 was used to calculate estimated travel

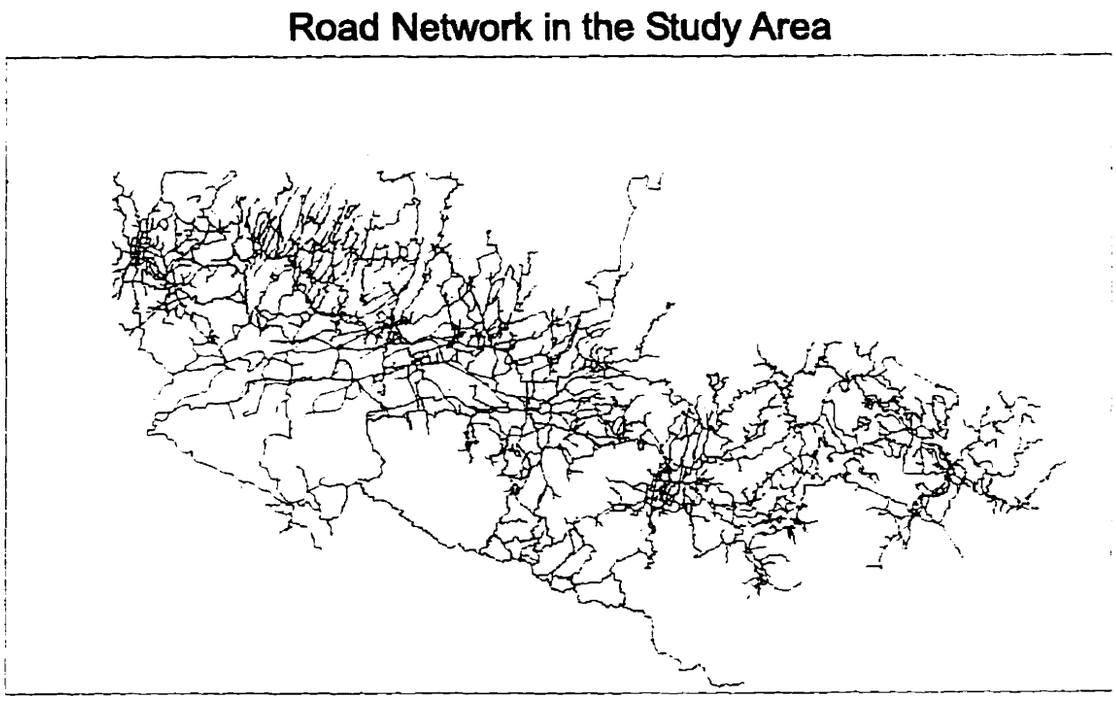


Figure 7.5: Road network in the study area.

times between supply and demand locations in the study area.

In order to estimate travel times, the different road types were weighted by an average speed. These speeds were consistent with those proposed by Entwisle *et al.* [1995] in estimating accessibility to family planning services in Thailand. Paved roads with two or more lanes were assigned a speed of 32 km/h. Paved roads with 1 lane and loose surface roads with two or more lanes were assigned a speed of 24 km/h while loose surface roads of only one lane were assigned a speed of 16 km/h. Finally, distances off the road network were assumed to be traversed at a speed of 8 km/h.

Since neither the grid-cells nor the service provider locations were constrained to lie exactly on the road network, the following procedure, illustrated in Figure 7.6, was used to calculate travel times. For each populated grid-cell and facility location, the straight line distance to, and the location of, the nearest point on the road network was found (indicated in Figure 7.6 as distance A and distance B respectively). A second straight line distance (distance C), between each grid-cell/facility pair, was also calculated. If distance C was found to be less than the total of the distance A plus distance B, then

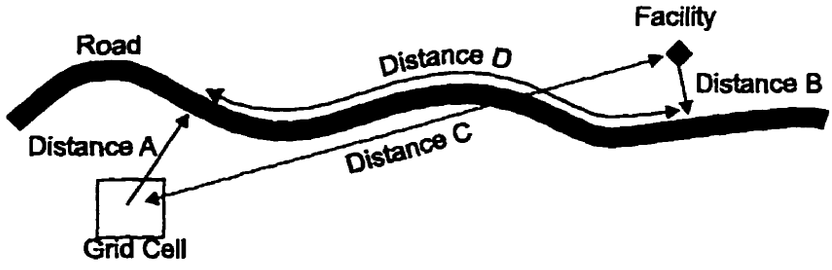


Figure 7.6: Example of calculation of travel times.

the distance C was used to calculate the travel time. Otherwise, the network distance between the locations on the road network nearest the grid-cell and the facility (distance D) was calculated and the total travel time was set to the appropriately weighted sum of distances A, B, and D.

7.2 Implications for Accessibility Modelling

As with any real-world data set, there are several issues relating to the accessibility measures and accessibility optimization models that require further examination. Specifically, the organization of the health care delivery system and the size of the data set both affect how the accessibility models are applied.

For this study area, the health care delivery system consists of three types of facilities, namely, hospitals, clinics, and health centres. The accessibility models should incorporate this organizational structure since different services are offered at the different types of facilities. For example, surgical sterilization is offered typically only at hospitals [CCSS, 1991], while the family planning methods available at health centres consist mainly of oral contraception, condoms, and IUDs [Ministerio de Salud, 1991]. To incorporate this information, these three types of facilities are arranged into a three-level successively-inclusive service hierarchy such as those described by Hodgson [1988] and Oppong [1992] and discussed in Section 2.3.1. Level A services, such as surgical sterilization, are offered only at hospitals. Level B services, such as injectables and diaphragms, are offered at both hospitals and clinics, while level C services, such as oral contraception, condoms, and IUDs are offered at all three facility types.

Although, the accessibility models discussed in this thesis have for the most part only considered a unitary level of service, they can be easily generalized to a three-level

service hierarchy. For evaluating accessibility, this implies calculating three interrelated accessibility values for each grid cell: level A accessibility to hospitals only, level B accessibility to hospitals and clinics, and, finally, level C accessibility to all three types of facilities. It is also possible to incorporate this service hierarchy into the accessibility optimization models using either the bottom-up or top-down approach discussed in Chapter 2.

A further and important issue relevant to the facility location accessibility optimization models is the set of candidate facility sites within the study area. The road intersections or nodes on the digitized road network were initially selected as the candidate facility sites. However, using all the nodes resulted in there being over 2400 candidate sites and this led to solution times of over twenty-four hours for some of the models[3]. To reduce these solution times, a reduced set of 611 candidate facility sites was selected from the road network shown in Figure 7.5. This set consisted of a random selection of 50% of the nodes in the San José region and 20% of the nodes in the remainder of the study area. This distribution was biased to account for the fact that the road network contained only major urban roads and omitted other streets.

Clearly, the set of candidate facility sites used for the location models directly influences the sites selected by the model, as well as the value of the optimization objectives. As well, heuristic, rather than exact, solution techniques for the facility location models may yield suboptimal solutions. As discussed in Chapter 6, the facility location models are *NP*-hard and, consequently, there are no known efficient solution techniques to find the optimal solution to problems of this type. Second, the main goal of applying these models is to illustrate their potential usefulness as exploratory tools for finding potential strategies for improving accessibility. Typically, a heuristic solution simply underestimates the potential improvement in efficiency and equity. To simplify the discussion that follows, these heuristic solutions will be called optimal solutions.

Despite these qualifications, the application of these models can provide a wealth of useful information about the nature of the current accessibility of fertile women to family planning services in the study area as well as producing solutions which can improve on both the equity and efficiency of the distribution of these services.

[3]Calculated on a 166 MHz Pentium with 32 Megabytes or memory.

7.3 Minimum Distance Accessibility

Recall that minimum distance accessibility is defined as the distance to the nearest facility. This distance is measured using both straight line distances and estimated travel times calculated using the procedure discussed previously. Further, three levels of accessibility are considered.

First, the accessibility of urban, suburban, and rural women in the fertile age cohort to all three levels of service is examined. Next, the distance-constrained p-median model is applied for two different planning scenarios: to determine the optimal facility configuration, and to determine the best locations for two new clinics and three new health centres.

7.3.1 Current Accessibility

For each of the 5028 populated grid-cells disaggregated from the distrito layer, two different values corresponding to one of the two distance measures, namely, straight line distances, and estimated road network travel times, were calculated for minimum distance accessibility for each of the three service levels discussed previously. These values were then weighted by the target population in each of the three populations groups, urban, suburban, and rural, in order to calculate the average and maximum distance (or travel time) to each service level for each population group as well as for the entire target population.

The overall results of the minimum distance accessibility calculations are shown in Table 7.1 with straight line distance calculations expressed in kilometres and travel time figures expressed in minutes. As expected, the average distance and maximum distance to level A services were much higher than to lower order services. The average weighted distance to level A services was approximately 4.5 km using straight line distance and 16.6 minutes using travel times compared to 1.7 km and 8.3 minutes for level C services. Similarly level A services were a maximum of 28.2 km or 115 minutes away whereas level C services were only 10.6 km and 57 minutes away. Note that the maximum distance for level A services for both the urban and rural population groups are approximately equal because the small urban population in the Distrito of Santiago, located at the tip of the southerly extremity of the study area west of San José is without a nearby hospital.

(a) Straight line distances

Service	Average distance (km)				Maximum distance (km)			
Level	Urban	Sub.	Rural	Avg.	Urban	Sub.	Rural	Max.
Level A	3.029	5.902	6.842	4.464	28.242	13.769	28.242	28.242
Level B	1.363	3.257	2.901	1.993	7.791	10.682	10.566	10.682
Level C	1.028	2.555	2.642	1.657	7.271	9.668	10.565	10.565

(b) Road network travel times

Service	Average travel time (min)				Maximum travel time (min)			
Level	Urban	Sub.	Rural	Avg.	Urban	Sub.	Rural	Max.
Level A	11.60	19.88	25.19	16.58	113.29	61.47	115.55	115.55
Level B	6.59	12.70	13.95	9.40	52.57	47.10	56.72	56.72
Level C	5.50	11.52	12.85	8.30	52.57	45.06	56.72	56.72

Table 7.1: Average and maximum minimum distance accessibility in the Central Valley of Costa Rica disaggregated by population group for *(a)* straight line distances and *(b)* road network travel times.

One interesting result of this analysis is the large difference in the average distance to the nearest facility among the population groups. For level C services, which are available at all facilities, women in urban areas were, on average, over twice as close as were women in suburban and rural areas using both straight line distances and travel times. Moreover, this pattern is evident for level A and level B services as well. While the difference in urban and rural accessibility is expected due to the difference in settlement patterns with urban areas having a much higher population density and, consequently a higher proximity to urban-based services, one surprising result from this analysis is the relatively poor accessibility of women in suburban areas to family planning services. As can be seen in Figure 7.2, the main concentration of this population is found in a ring around the city of San José. Although this region includes a large number of facilities, the accessibility of this population group is approximately the same as in rural areas which have a much more dispersed population pattern.

This pattern of differential accessibility is also evident when examining the distribution of accessibility within the population. Table 7.2 shows the number of women in each population group tabulated by their accessibility to level C services, while the percentage of each population group by their level of accessibility is presented in Figure 7.7. The distribution of women, using straight line distances and travel times, is broadly similar. For level C services, over 90% of all women in urban areas are within

Access	Straight line distances				Road Network Travel times				Access
(km)	Urban	Sub.	Rural	Total	Urban	Sub.	Rural	Total	(min)
< 1	165 719	6777	18 221	190 717	160 228	6664	20 223	187 115	< 5
1–2	92 730	8769	41 054	142 553	95 568	10 670	46 104	152 342	5–10
2–3	19 153	5863	41 088	66 104	24 365	10 171	59 986	94 522	10–20
3–5	5192	7400	38 077	50 669	2477	3311	18 829	24 617	20–30
5–8	313	3352	12 379	16 044	452	1501	5836	7789	30–45
> 8	0	158	449	607	17	2	290	309	> 45

Table 7.2: Distribution of population by existing minimum distance accessibility to level C services in the Central Valley of Costa Rica.

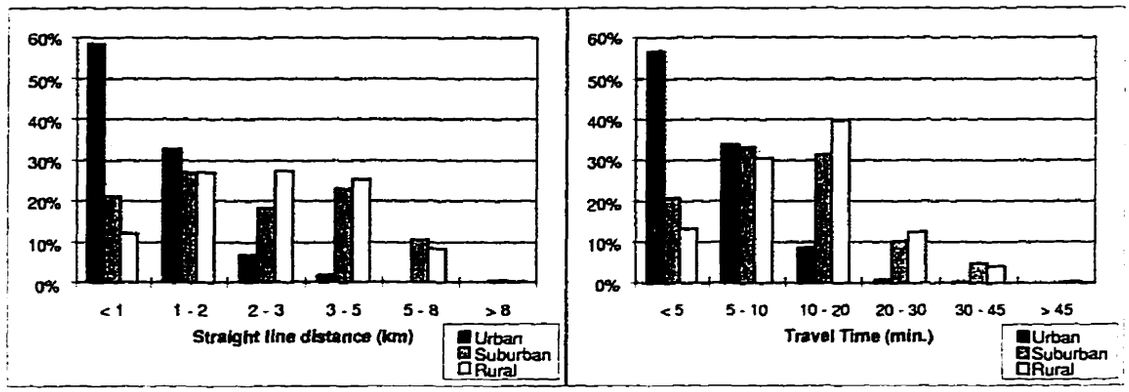


Figure 7.7: Percentage distribution of population by existing minimum distance accessibility to level C services.

2 km or 10 minutes travel time of the nearest facility. This compares to approximately 50% of suburban women and 45% of women in rural areas.

One of the most useful aspects of using a population grid is the ability to display the spatial distribution of accessibility or the accessibility "surface." Figure 7.8 illustrates the accessibility to level C services for straight line distances and travel times. Note that the grid-cells are shaded according to the accessibility classifications used in Table 7.2. Both surfaces exhibit the same basic pattern of accessibility, with areas of high accessibility concentrated in the major cities and areas of low accessibility concentrated mainly around eastern and southern fringes of San José and in pockets in the western regions of the Central Valley. Using straight line distances result in a fairly smooth accessibility surface, with gradually changing levels of accessibility between grid-cells.

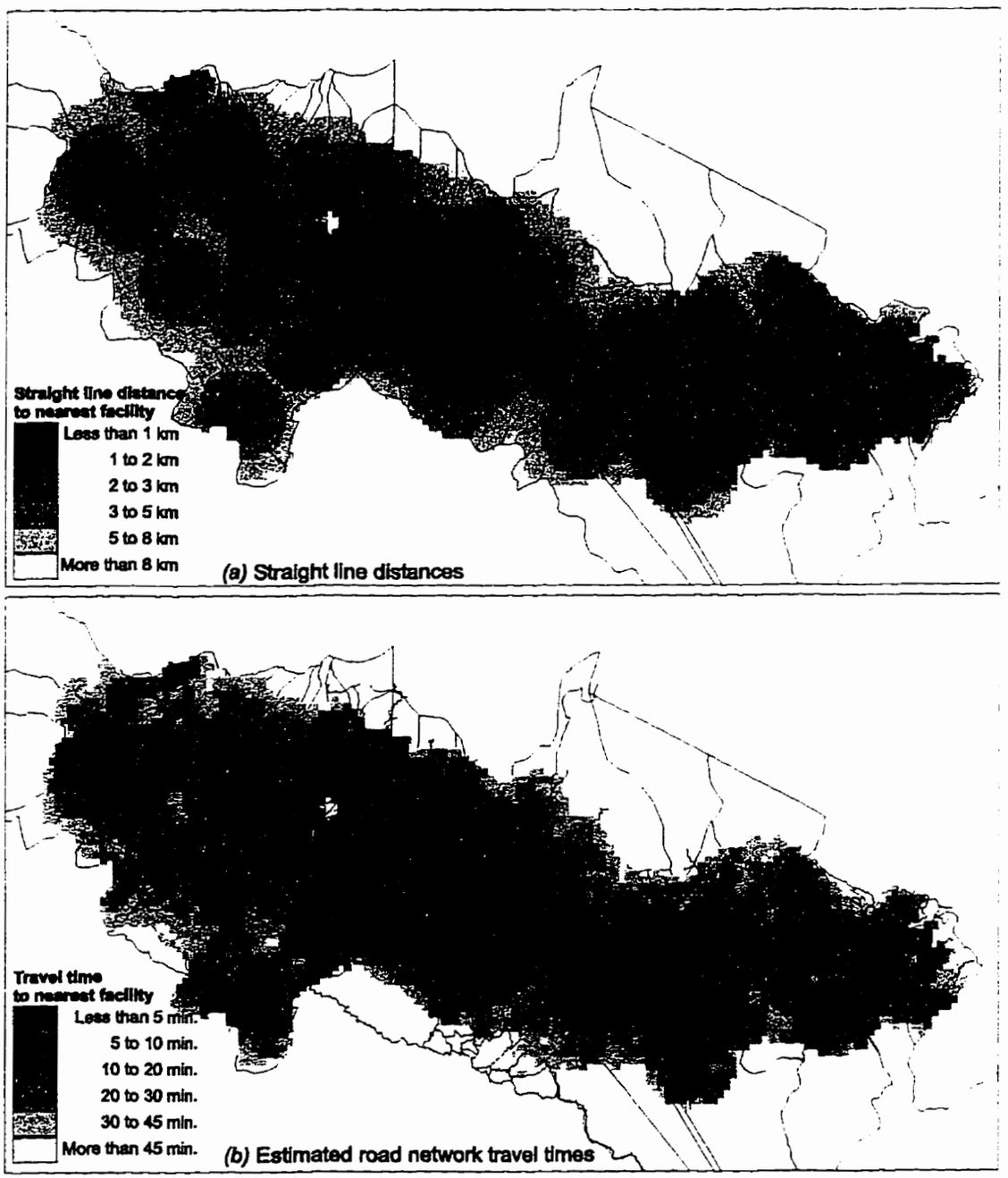


Figure 7.8: Existing minimum distance level C accessibility in the Central Valley of Costa Rica measured *(a)* using straight line distances and *(b)* using estimated road network travel times.

However, accessibility calculated using road network travel times does not exhibit this same smooth change. Instead, there is a much more complicated pattern of accessibility, with, as one would expect, areas of greater accessibility generally following the road pattern.

7.3.2 Accessibility Optimization

The previous section described the minimum distance accessibility of the target population to the health care delivery system in 1992. In particular, it noted that there was a large discrepancy in average accessibility between urban areas and suburban/rural areas. This raises two questions about the current accessibility of the system. First, how much of this differential accessibility is caused by the settlement pattern? Urban areas are much more densely populated while the population in rural areas is more widely dispersed. It may not be possible to increase significantly rural accessibility without dramatically reducing the overall accessibility to the target population. Second, how would the addition of new facilities affect the overall level of accessibility? Where could the facilities be located and how would these new facilities affect the accessibility of the population groups?

In order to help answer these questions, a three-level hierarchical distance-constrained p-median model was applied to this data set for two different optimization scenarios using both straight line distances and travel times. The first scenario, "full optimization," determines an optimal facility configuration while the second scenario adds two new clinics and three new health centres to the existing set of facilities. The Interchange heuristic is used to solve the p-median problem using the existing system as the original facility configuration. If the existing system violates the distance constraints then the set-covering problem is applied in order to determine a feasible initial solution. If this feasible solution has fewer than p facilities, the Add heuristic is then applied to find an initial solution with the required number of facilities.

A bottom-up approach is used to solve the hierarchical problem. Although Hodgson [1984] demonstrated that a simultaneous approach performed somewhat better than a bottom-up approach and much better than a top-down approach, the simultaneous method requires that each service level be assigned a weight reflecting their relative usage requiring information on the the usage of each level of service at each facility. Since this information was not available, a bottom-up approach is taken in this

problem. Facilities offering level C services (all facilities) are located first, these selected sites then become the candidate facility sites for locating level B (hospitals and clinics) facilities and these are then used to locate level A facilities (hospitals).

The p-median model is also used to examine the trade-off between efficiency (average distance) and equity (maximum distance) by adjusting the maximum distance constraints. Two different solutions are identified in each scenario for both distance measures. First, the efficiency solution is generated with no maximum distance constraints applied. Second, the equity solution is determined by successively reducing the maximum distance constraint for each level to find the facility configuration with the smallest maximum distance. This is done first for level C services, then for level B, and finally for level A services.

Full Optimization Scenario

In the full optimization scenario, a new optimal facility configuration is determined assuming that there were no fixed facilities, *i.e.*, the p-median model finds an optimal configuration of 14 hospitals, 58 clinics, and 48 health centres so as to minimize the average distance of 15-49 year old women to the nearest facility. This is done using both straight line distances and road network travel times to find both the efficiency and equity solutions. Thus, four optimal solutions are calculated: the straight line distance efficiency solution, the straight line distance equity solution, the travel time efficiency solution, and the travel time equity solution. On a 166 MHz Pentium PC running Windows NT, the execution time to calculate the efficiency solution was 2 hours and 33 minutes while the execution time for computing the equity solution was 3 hours and 2 minutes using straight line distances were used. When road network travel times were used these solutions required 1 hour and 18 minutes and 1 hour and 19 minutes respectively[4]. The accessibility indicators of these solutions and the percentage change in travel time and distance from the existing system are shown in Table 7.3.

Several interesting trends are apparent for these results. For level A services (hospitals), all the solutions reduced the average distance or travel time by 15% to 23% except for the travel time equity solution. This solution actually increased the average travel time by 5% while reducing the maximum travel time by 43% from 115 minutes to 67 minutes. The straight line distance equity solution reduced the maximum distance

[4]The difference in execution times is due to the fact that the travel times were pre-calculated.

(a) Straight line distances

Service Level	Average distance (km) Urban	Sub.	Rural	Avg.	Maximum distance (km) Urban	Sub.	Rural	Max.
Efficiency solution								
Level A	2.320	4.360	5.568	3.514	13.162	11.262	17.843	17.843
	-23.39%	-26.12%	-18.63%	-21.28%	-53.39%	-18.21%	-36.82%	-36.82%
Level B	1.102	1.716	2.472	1.588	7.239	6.054	10.565	10.565
	-19.20%	-47.31%	-14.78%	-20.30%	-7.08%	-43.32%	-0.01%	-1.09%
Level C	0.862	1.351	1.935	1.243	5.440	5.176	8.777	8.777
	-16.22%	-47.13%	-26.77%	-24.97%	-25.18%	-46.46%	-16.93%	-16.93%
Equity solution								
Level A	2.736	4.359	5.684	3.804	13.333	13.854	14.173	14.173
	-9.68%	-26.13%	-16.93%	-14.79%	-52.79%	+0.62%	-49.82%	-49.82%
Level B	1.122	1.723	2.539	1.623	5.855	6.384	7.747	7.747
	-17.73%	-47.11%	-12.47%	-18.57%	-24.85%	-40.23%	-26.68%	-27.47%
Level C	0.880	1.374	1.913	1.249	5.535	5.455	7.747	7.747
	-14.42%	-46.21%	-27.60%	-24.62%	-23.88%	-43.58%	-26.67%	-26.67%

(b) Road network travel times

Service Level	Average travel time (min) Urban	Sub.	Rural	Avg.	Maximum travel time (min) Urban	Sub.	Rural	Max.
Efficiency solution								
Level A	8.18	16.08	20.63	12.76	53.13	55.42	103.02	103.02
	-29.49%	-19.12%	-18.11%	-23.03%	-53.10%	-9.85%	-10.84%	-10.84%
Level B	4.98	7.90	11.78	7.39	42.63	40.51	67.39	67.39
	-24.40%	-37.78%	-15.51%	-21.38%	-18.91%	-13.98%	+18.80%	+18.80%
Level C	4.32	6.77	9.85	6.28	42.44	33.67	67.39	67.39
	-21.48%	-41.22%	-23.34%	-24.31%	-19.26%	-25.28%	+18.80%	+18.80%
Equity solution								
Level A	13.73	22.45	23.20	17.40	64.76	63.97	65.94	65.94
	+18.34%	+12.96%	-7.89%	+4.98%	-42.84%	+4.07%	-42.93%	-42.93%
Level B	5.22	9.10	11.91	7.66	42.63	43.29	46.24	46.24
	-20.75%	-28.29%	-14.60%	-18.50%	-18.91%	-8.08%	-18.48%	-18.48%
Level C	4.36	6.91	9.89	6.33	42.44	34.97	46.24	46.24
	-20.71%	-40.02%	-22.98%	-23.70%	-19.26%	-22.40%	-18.48%	-18.48%

Table 7.3: Accessibility indicators and percentage change from existing values for the maximum efficiency and maximum equity solutions of the full optimization scenario.

from 28.2 km to 14.2 km, an improvement of nearly 50%. Further, the two efficiency solutions reduced maximum straight line distance by 37% and the maximum travel time by 11%. One consistent feature of all the solutions was a large reduction, between 43% and 53%, in maximum distance or travel time for the urban population. Other than this feature, there were no consistent trends in the change in differential accessibility to level A services among the three population groups.

With respect to the accessibility of level B services (offered at hospitals and clinics), the efficiency solutions reduced the average distance or travel time by about 20% with a slight decrease in the maximum straight line distance and an almost 20% increase in maximum travel time. The increase in maximum travel time likely results from several sparsely populated rural areas distant for a road not having a facility located near them in the optimal solutions. However, the equity solutions reduced both the average distance and travel time by 18.5% while also decreasing the maximums by 18.5% to 27.5%. These numbers mask a consistent pattern in the change in differential accessibility. For every solution, the suburban population experienced a dramatic improvement in accessibility. For example, the average distance was reduced by 47% from 3.26 km to 1.72 km for both straight line distance solutions whereas the average travel time for the suburban population was reduced from nearly 13 minutes to between 8 and 9 minutes. Urban areas experienced the second largest reduction while the improvement in rural areas was less than the average.

A similar trend was evident for level C minimum distance accessibility. For all solutions, the average distance was reduced by almost 25%. In addition, the equity solutions reduced the maximum distance by between 18% and 27% while the straight line distance efficiency solution decreased it by 17%. As with level B services, the travel time efficiency solution increased the maximum travel time by more than 10 minutes as it did not locate facilities near several isolated rural areas. Again, the suburban population received the largest increase in accessibility. This increase ranged from 40% to 47%. The accessibility improvements in rural areas ranged from 23% to 25%, and in urban areas from 14% to 21%.

The change in differential accessibility can be even more clearly observed by examining the distribution of accessibility among the population groups. Table 7.4 shows the percentage of the target population than can be considered to have "good" accessibil-

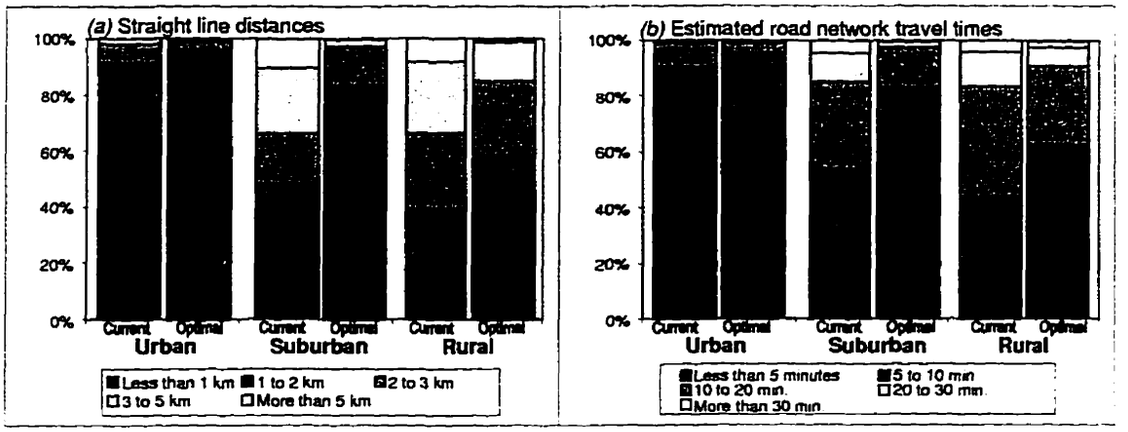


Figure 7.9: Distribution of population by accessibility classes for *(a)* straight line distance and *(b)* estimated road network travel times for the existing system (current) and the efficiency solutions (optimal).

ity[5]. For the target population residing in a urban area, the optimal solutions increased this percentage by about 5%. However, the most dramatic increase, from approximately 50% to 80%, occurred for the suburban population. For rural areas, the increase was from about 40% to 60%. This same pattern can be seen in Figure 7.9 which shows the distribution of population, by accessibility classes, for the current system and the two efficiency solutions.

It is also possible to illustrate the pattern of accessibility changes spatially. Figure 7.10 depicts two surfaces showing the change in existing level C accessibility resulting from the efficiency solutions for both straight line distances and travel times. The figure also shows the current and optimal facility locations for these two cases. Both surfaces exhibit the same basic pattern of accessibility change, with the areas of greatest increase around the eastern boundary of the San José metropolitan area and scattered in the western regions of the Central Valley. The areas of decrease were mainly in the eastern sections of the study area. Further, the optimal facility locations were more dispersed than the existing facilities.

These results clearly indicate that an optimal configuration of facilities can improve minimum distance accessibility considerably while also reducing the maximum distance. Although the results for accessibility to hospitals are less noticeable, the most

[5]Good accessibility is defined as being within 2 km or 10 minutes of a facility.

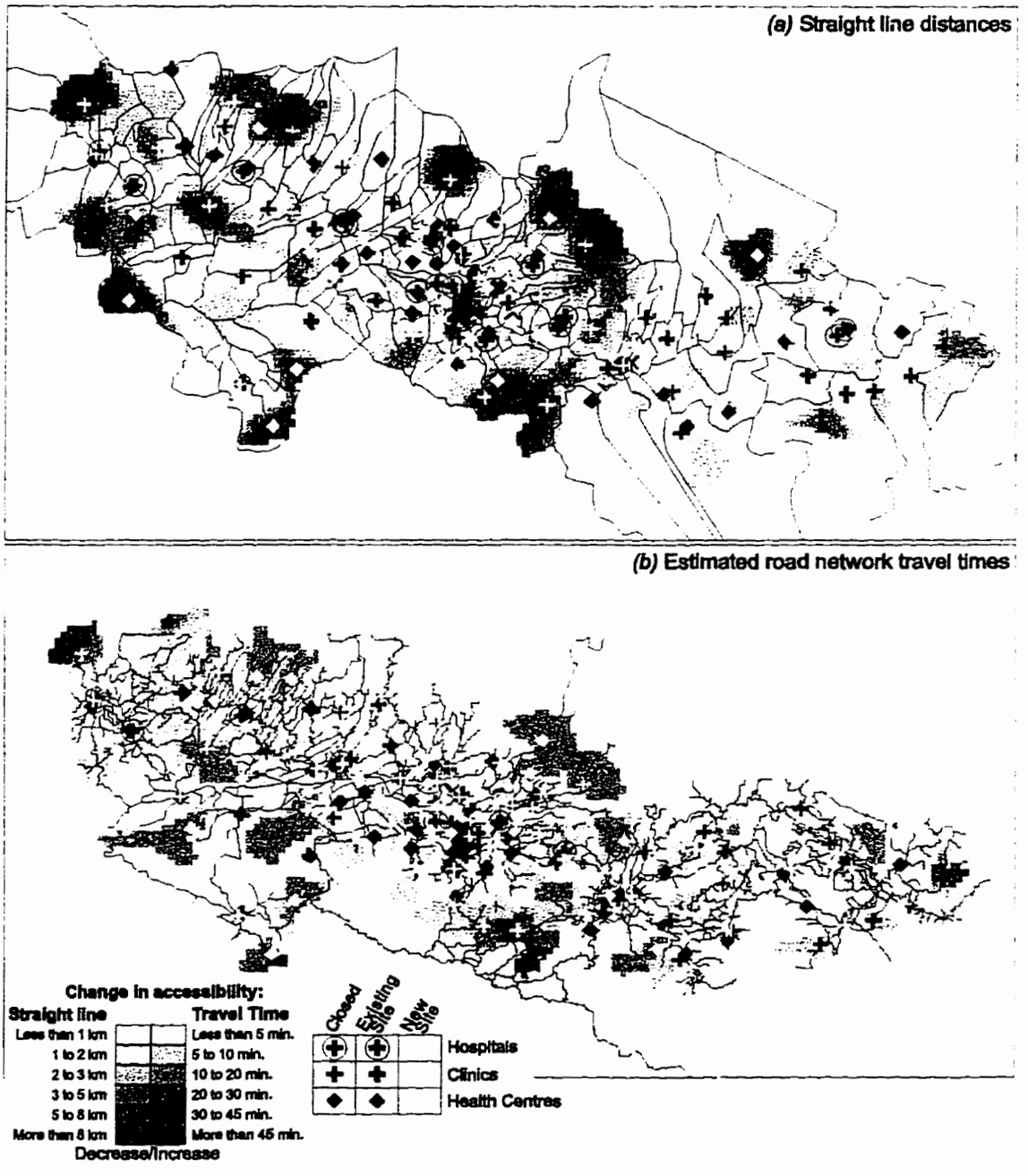


Figure 7.10: Change in level C minimum distance accessibility and facility locations for maximum efficiency solutions for the full optimization scenario.

(a) Straight line distances, Percentage < 2 km

Scenario	Solution	Urban	Sub.	Rural	Total
–	Current system	91.29%	48.10%	39.19%	71.41%
Full	Efficiency solution	96.48%	83.39%	58.29%	83.19%
	Equity solution	96.48%	82.70%	59.52%	83.54%
Additional	Efficiency solution	91.70%	61.49%	41.65%	73.39%
	Equity solution	91.30%	56.81%	40.57%	72.47%

(b) Travel times, Percentage < 10 minutes

–	Current System	90.35%	53.63%	43.85%	72.74%
Full	Efficiency solution	92.38%	61.64%	46.66%	75.43%
	Equity solution	90.63%	60.44%	45.85%	74.02%
Additional	Efficiency solution	92.38%	61.64%	46.66%	75.43%
	Equity solution	90.63%	60.44%	0.57%	72.47%

Table 7.4: Percentage of women *(a)* less than 2 km, *(b)* less than 10 minutes from a facility for existing system and for optimal solutions.

dramatic increases occur for the suburban population mainly located at the outer boundaries of San José. One possible explanation for the relatively low accessibility of suburban areas is that these areas have experienced recent population growth, while there is evidently a lag in the decentralization of services in response to the changing pattern of demand. This disparity is probably exacerbated in reality as the estimated population counts did not include the effects of in-migration which would likely be important in the suburban area.

Additional Five Facilities Optimization Scenario

Relocating nearly every facility is clearly not a feasible option to improve accessibility. Instead, a more reasonable strategy is to open additional facilities in areas of greatest need. The second optimization scenario directly applies this strategy by locating two additional clinics and three additional health centres while keeping all existing facilities in their current locations. Again, efficiency and equity solutions were calculated using both straight line distances and road network travel times. For this scenario, the execution times were 190 seconds for the efficiency solution using straight lines and 147 seconds for the equity solution. Using the network travel times, the solution time for the efficiency solution was 100 seconds and 71 seconds for the equity solution. The

(a) Straight line distances

Service Level	Average distance (km) Urban	Sub.	Rural	Avg.	Maximum distance (km) Urban	Sub.	Rural	Max.
Efficiency solution								
Level B	1.358	2.393	2.838	1.909	7.318	8.620	10.566	10.566
	-0.08%	-7.49%	-2.04%	-1.84%	0.00%	-14.98%	-7.78%	-8.78%
Level C	1.016	1.856	2.488	1.551	7.271	8.620	10.565	10.565
	-0.97%	-23.40%	-6.34%	-6.13%	0.00%	-35.59%	-12.34%	-12.34%
Equity solution								
Level B	1.362	3.023	2.854	1.961	7.791	9.082	9.744	9.744
	-0.08%	-7.20%	-1.64%	-1.62%	0.00%	-14.98%	-7.78%	-8.78%
Level C	1.027	1.969	2.509	1.573	7.271	8.617	8.617	8.617
	-0.11%	-22.91%	-5.06%	-5.10%	0.00%	-10.87%	-18.44%	-18.44%

(b) Road network travel times

Service Level	Average travel time (min) Urban	Sub.	Rural	Avg.	Maximum travel time (min) Urban	Sub.	Rural	Max.
Efficiency solution								
Level B	6.42	11.37	13.76	9.14	52.57	47.10	56.72	56.72
	-2.50%	-10.41%	-1.37%	-2.70%	0.00%	0.00%	0.00%	0.00%
Level C	5.13	9.86	12.50	7.85	52.57	45.06	56.72	56.72
	-6.75%	-14.39%	-2.69%	-5.45%	0.00%	0.00%	0.00%	0.00%
Equity solution								
Level B	6.59	12.70	13.75	9.34	52.57	47.10	53.81	53.81
	0.00%	0.00%	-1.39%	-0.54%	0.00%	0.00%	-5.13%	-5.13%
Level C	5.47	10.39	12.45	8.07	42.92	45.06	51.17	51.17
	-0.61%	-9.77%	-3.07%	-2.73%	-18.36%	0.00%	-9.79%	-9.79%

Table 7.5: Accessibility indicators and percentage change from existing values for the maximum efficiency and maximum equity solutions for the add five optimization scenario.

accessibility indicators of these solutions and the percentage change from the existing system are shown in Table 7.5. Level A results are omitted as the configuration of hospitals was unchanged.

In comparison to the full optimization scenario, the accessibility gains achieved by adding the 5 facilities were much more modest and ranged from a minimum of 0.5% (for level B services with the travel time equity solution) to a maximum of 6.1% (for level C services for the straight line distance efficiency solution). Level C accessibility improved more than level B accessibility and the efficiency solutions had a greater impact on average distance than the equity solution. This trend was expected since 5 new facilities offering level C services were located compared to only 2 new level B facilities. With the exception of the travel time efficiency solution, which did not affect the maximum travel time, all of the solutions reduced the maximum distance by 5.1% to 18.4%, with level C services decreasing more than level B services.

As with the full optimization scenario, suburban population showed the largest increase in potential accessibility while rural areas exhibited a smaller improvement and, except for the travel time efficiency solution, urban populations experienced only a marginal change in accessibility. One explanation for these changes in accessibility is that only 7% of the population is suburban, so that a small change in the facility locations may disproportionately affect this group. The suburban population also received the largest accessibility increases in the full optimization scenario solutions. An explanation of the large increases in suburban accessibility can be found by examining where the new facilities are located. Figure 7.11 shows the new facilities and the change in the potential accessibility surface for the two efficiency solutions. For the travel time efficiency solution, all the new facilities are located in a ring around central San José while the straight line distance solution located three new facilities, including both clinics, in this region. The straight line distance equity solution (not shown) exhibited a similar pattern to the efficiency solution while the travel time equity solution (also not shown) placed only two facilities in the eastern fringes of San José and three facilities in the western Central Valley.

In conclusion, these two scenarios indicate that, in terms of minimum distance accessibility, existing services are sub-optimally located, particularly for the suburban population. In particular, the fringe of the San José region is relatively under-supplied as are areas in the western Central Valley. Compared to an optimal system, the average distance to the nearest facility is between 31% and 33% greater for the existing system.

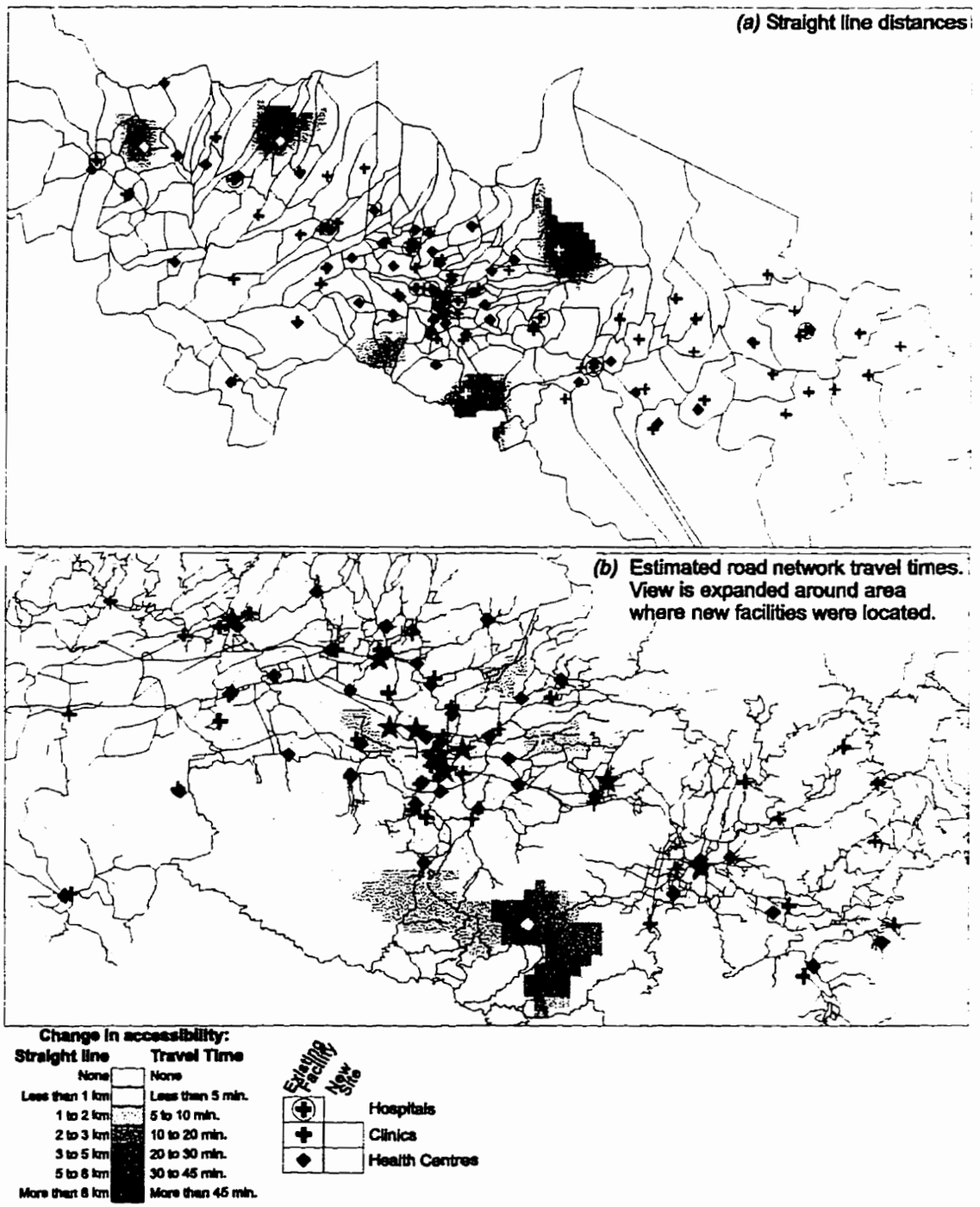


Figure 7.11: Change in level C minimum distance accessibility and facility locations for maximum efficiency solutions for add five scenario.

This figure is remarkably similar to the values of 26% reported by Oppong and Hodgson [1994] in Ghana and 30% produced by Ayeni *et al.* [1987] in Nigeria using similar analyses.

Hodgson [1988] points out some of the limitations with minimum distance accessibility. These shortcomings include the assumption that all residents at a given location attend the nearest facility, and that accessibility varies linearly with distance. Further, minimum distance accessibility does not consider the resource availability at a facility. The Joseph and Bantock [1982] accessibility measure, applied in the next section, overcomes these limitations.

7.4 Joseph and Bantock Accessibility

This section applies an alternative measure of potential geographic health care accessibility to the study area, namely the Joseph and Bantock [1982] accessibility measure (abbreviated as J&B accessibility). This measure was discussed in Example 3 of Chapter 3 and in Chapter 6. As before, the target population is split into urban, suburban, and rural components and a three-level successively-inclusive model of accessibility is used.

As currently formulated, the J&B accessibility measure only considers a single level of service. Using this measure, the accessibility of area i is defined as

$$A_i = \sum_j (S_j/C_j) \exp(-\beta D_{ij}) \tag{7.1}$$

where S_j is the resource availability at facility j and

$$C_j = \sum_i P_i \exp(-\beta D_{ij})$$

is the potential demand on facility j. There are two possible approaches to adapt this measure to consider hierarchical service levels. The first approach would be to use a service-level specific decay parameter, β_k, and the level-specific resource availability at a given facility, S_{jk}. However, as noted earlier, it can be difficult to obtain resource availability data disaggregated by service levels. An alternate approach is to use facility-type specific decay parameters so that only the total resource availability at a given facility is required. Using this approach, the hierarchical J&B accessibility measure for sub-area i

to services of level k can be defined as

$$A_i^k = \sum_{\ell=1}^{k} \sum_{j \in \mathcal{F}_k} (S_j / C_j) \exp\left[-\beta_{h(i),\ell} D_{ij}\right] \tag{7.2}$$

where $C_j = \sum_i P_i \exp\left[-\beta_{h(i),\ell} D_{ij}\right]$, $\mathcal{F}_k$ is the the set of facilities of type k, $h(i)$ is the type of population sub-group in sub-area i[6], and β_{hk} is the distance decay parameter for population sub-group h to facility type k.

In order to apply this model, nine distance decay parameter values are required: one for each combination of population sub-group (urban, suburban, and rural) and facility type (hospitals, clinics, and health centres). The next section describes the calibration procedure used to obtain estimates of these values. Then, the existing J&B potential accessibility of the population in the study area is discussed. The two accessibility optimization models developed in the previous chapter for this accessibility measure, the facility location subproblem (FLS) and the resource allocation subproblem (RAS), are then applied in turn.

7.4.1 Calibration

The J&B accessibility measure requires the specification of distance decay parameters that describe specified components of spatial behaviour. Parameter estimates were obtained using data from the 1992 Costa Rican Reproductive Health Survey (Spanish Acronym ESR) conducted by the CCSS with assistance from the US Center for Disease Control and Prevention (CDC). In this survey, a nationally representative sample of 3618 women in the fertile age cohort, 15-49, was selected from 188 segmentos or census tracts from the 1984 Costa Rican Population Census [CCSS, 1994]. One notable feature of the ESR was that both the census tract and the health facility attended were coded in the survey. Dr. Luis Rosero of the University of Costa Rica subsequently geocoded a database of the survey clusters locations and of the service provider locations. Observations for 1203 women from 90 survey clusters in the study area were obtained from him. The survey was used to identify which (if any) public health care facility was most recently attended to obtain family planning services. Thus, the survey provided information on the spatial distribution of trip patterns for attenders as well as the locations

[6]The sub-areas are defined so that each sub-area is composed of only one population sub-group. With reference to population grid, this implies that grid-cells may be represented by up to three sub-areas, one each for the urban, suburban, and rural populations.

Population Group	Last Facility Attended				
	Hospital	Clinic	Centre	None	Total
Urban	137	92	56	242	527
Suburban	38	32	35	50	155
Rural	131	90	137	163	521
Total	306	214	228	455	1203

Table 7.6: Attendance patterns from the survey data.

of the non-attenders. The pattern of attendance and non-attendance is summarized in Table 7.6.

In order to estimate the distance decay parameters, it is necessary to know the distance or travel time of each woman to each of the facilities. The straight line distances and estimated road network travel times were calculated between each of the 90 survey clusters and the 120 facilities. Each woman was assumed to be located at the survey cluster so that the distances and travel times for each woman were those of her survey cluster. These data were used to calibrate a multinomial logit (MNL) model, discussed in Section 3.3 and defined by equation (3.40), to obtain estimates of the distance decay parameters for the accessibility measure. For each individual, the feasible choice set was composed of the null alternative, *i.e.*, non-attendance, and each of the facilities, so that there were a total of 121 feasible alternatives. Further, the survey data were partitioned into the three population groups, urban ($h = 1$), suburban ($h = 2$), and rural ($h = 3$) and the facilities were denoted by $k = 1$ for hospitals, $k = 2$ for clinics, and $k = 3$ for health centres. Following Example 3 in Chapter 3, the measured attractiveness, V_{ij}, of individual i in population group h to alternative j was defined as

$$V_{ij} = \begin{cases} H_h & \text{if } j = 0, \text{ the null alternative,} \\ \ln S_j - \ln C_j - \sum_{k=1}^{3} \beta_{hk} D_{ij}^k & \text{otherwise,} \end{cases} \tag{7.3}$$

where H_h is the attractiveness of non-attendance of a person in population group h and $D_{ij}^k = D_{ij}$, if facility j is of type k, and equals zero otherwise. All model parameters were specific to each population group, that is, there were no common model parameters. Consequently, the MNL was calibrated, using maximum likelihood estimation[7],

[7]See, for example, [Ben-Akiva and Lerman, 1985] for a discussion of maximum likelihood estimation

separately for each population group.

To fit this MNL model, information was required on the potential demand, C_j, on each of the facilities. Since this value is dependent upon the values of the distance decay parameters that describe facility use, an iterative approach was adopted. The initial decay parameters were arbitrarily set to 0.5 (per kilometre) for straight line distances, and 0.16 (per minute) for travel times and the potential demand on each facility was evaluated using these initial parameters. Next, the maximum likelihood estimates for the model parameters were calculated for each population group. Each facility's potential demand was then re-assessed using the new parameter estimates and the MNL model was re-calibrated. This process was repeated until the parameter estimates were stable[8].

Table 7.7 gives the final results for the calibration procedure for the three population groups using both straight line distances and travel times. As indicated by the χ^2 statistic (for 4 degrees of freedom) each of the models is statistically significant at the 99.5% confidence level. This indicates that, unsurprisingly, the attendance patterns were not random and that distance does have an effect on facility choice. This is also confirmed by ρ^2, an informal goodness-of-fit measure. In addition, the t-statistics indicate that each parameter is significant at the 99.5% confidence level. As expected, the distance decay parameters for the hospitals, $\beta_{\bullet 1}$, were smaller than the parameters for clinics, $\beta_{\bullet 2}$, or health centres, $\beta_{\bullet 3}$. This indicates that people are willing to travel further distances to attend a hospital than to attend either clinics or health centres. Further, the decay parameters for the urban population were larger than the suburban or rural parameters, indicating that people in urban areas were more sensitive to the effects of distance. Based on these statistics, the straight line distances had the best fit for the urban population while the travel time model was best for rural areas. There was no significant difference between the two models for the suburban population.

The next section outlines the results obtained from evaluating accessibility in the study area using the hierarchical J&B accessibility measure and the parameter values developed in this section.

---

of multinomial logit models

[8]In fact, the estimates were not very sensitive to potential demand and the procedure terminated on the third iteration.

Variable	Straight line distances (km) Coefficient Estimate	Asymptotic Standard Error	*t*-Statistic	Travel time (min) Coefficient Estimate	Asymptotic Standard Error	*t*-Statistic
Urban						
H_1	-1.486	0.187	-7.9	-1.867	0.203	-9.2
β_{11}	0.1230	0.0129	9.5	0.0531	0.0053	10.0
β_{12}	0.8421	0.0559	15.1	0.2530	0.0146	17.3
β_{13}	1.0204	0.0706	14.5	0.3117	0.0187	16.7
χ^2		2632			2611	
ρ^2		0.518			0.514	
Suburban						
H_2	-1.662	0.304	-5.5	-1.764	0.302	-5.8
β_{21}	0.1016	0.0191	5.3	0.0389	0.0071	5.5
β_{22}	0.4974	0.0508	9.8	0.1710	0.0167	10.2
β_{23}	0.5907	0.0617	9.6	0.2018	0.0199	10.1
χ^2		686.4			684.9	
ρ^2		0.459			0.458	
Rural						
H_3	-1.927	0.169	-11.4	-2.235	0.175	-12.8
β_{31}	0.1016	0.0100	10.1	0.0369	0.0033	11.1
β_{32}	0.5011	0.0271	18.5	0.1457	0.0073	20.1
β_{33}	0.5842	0.0327	17.9	0.1726	0.0091	19.0
χ^2		2327			2345	
ρ^2		0.463			0.467	

Table 7.7: Parameter estimates for final MNL model calibration.

7.4.2 Current Accessibility

The accessibility of the three population groups in the 5028 populated grid cells was evaluated using the distance decay parameters obtained from calibrating the MNL model for the three services levels using straight line distances and estimated road network travel times. The resource availability at each facility was given by the number of hours of family planning service available in 1992. This amounted to a total of 16 015 hours at the 14 hospitals, 64 850 hours at the 58 clinics, and 59 021 hours at the 48 health centres. The accessibility of each grid-cell was calculated as the population-weighted average of the accessibility levels for each of the three population groups in that grid-cell. Also, the accessibility indicators discussed in the previous chapter, the average satisfaction and the coefficient of variation of the three service levels were calculated for each population group and for the entire target population.

Straight line distances

Service	Average satisfaction				Coefficient of variation			
Level	Urban	Sub.	Rural	All	Urban	Sub.	Rural	All
Level A	-3.3423	-3.4742	-3.6760	-3.4596	0.3737	0.2364	0.4083	0.3776
Level B	-1.9250	-1.9460	-1.9667	-1.9400	0.5359	0.7283	0.6349	0.5845
Level C	-1.3749	-1.4405	-1.4026	-1.3884	0.5519	1.0676	0.7059	0.6521
Road network travel times								
Level A	-3.4130	-3.4749	-3.6474	-3.4933	0.4216	0.3028	0.4850	0.4364
Level B	-2.0295	-2.1515	-2.0351	-2.0398	0.6360	0.7491	1.0798	0.8137
Level C	-1.5315	-1.6703	-1.4884	-1.5272	0.6034	1.0188	1.1482	0.8485

Table 7.8: Existing J&B accessibility indicators for the Central Valley of Costa Rica.

The results of evaluating J&B accessibility for the study area are presented in Table 7.8. Note that for each level, the average accessibility is equal to Q/P_T so that average accessibility equals 0.03431 for level A services, 0.1733 for level B services, and 0.2997 for level C services. Therefore, the coefficient of variation is calculated by dividing the standard deviation by the respective average accessibility. For example, the standard deviation of accessibility of level C services is 0.1954 so that the corresponding coefficient of variation is 0.6521. Since the average accessibility is a constant, only

the coefficients of variation are reported. The standard deviation can be obtained by multiplying by the appropriate average accessibility value.

Using $\ln(Q/P_T)$ to compute the maximum possible value of the average satisfaction for a total population of 466 694, the upper bound for average satisfaction is -3.372 for level A services (hospitals only), -1.752 for level B services (hospitals and clinics), and -1.205 for level C services (all facilities). The average satisfaction levels for the target population were all below these values, with the indicators from the travel time model being slightly lower than from the straight line distance model. Further, the coefficient of variation was the lowest for hospitals (level A services) and the highest for all facilities (level C services). Even though there were only 14 hospitals within the study area, the coefficient of variation to them was the smallest. The explanation for this is that the distance decay parameter for hospitals was smaller than for other facility types indicating that the attractiveness of hospitals decayed slowly with distance resulting in a relatively even distribution of accessibility and, consequently, a smaller coefficient of variation.

The overall accessibility indicators hide large differences both in the average satisfaction and in the coefficient of variation among the population groups. For level A services, the rural population consistently had the lowest average satisfaction and the highest coefficient of variation while urban areas had the highest average satisfaction using straight line distances and suburban areas had the highest with the travel time model. In fact, using straight line distances, the average urban satisfaction to level A services of -3.342 slightly exceeds the upper bound on average satisfaction, -3.372, of the entire target population. In addition, the coefficient of variation for level A services was lower for suburban areas than for urban areas in both cases. The low value of average satisfaction for urban areas when travel times are considered is a result of several urban areas in the western part of the study area being 110 minutes away from a hospital, while all the the suburban population is within 65 minutes of a hospital. The decay parameter is also higher for the urban population, which further magnifies these differences. Consequently, these urban areas have very poor level A accessibility thus lowering the average satisfaction and increasing the coefficient of variation. Both these measures are sensitive to extreme values. In fact, the average satisfaction indicator is extremely sensitive to low accessibility values and approaches $-\infty$ if any area is considered inaccessible.

For level C services, the suburban population had the lowest satisfaction with a

high coefficient of variation, indicating that accessibility was unevenly distributed and there were areas with very low accessibility. A surprising result was that rural areas had the highest average satisfaction when using travel times although the coefficient of variation was very high. This probably reflects the presence of several medium-sized facilities with low potential demand in areas with a sparse road network. The population located in these areas would have very high accessibility which would raise the average satisfaction as well as increasing the coefficient of variation.

(a) Straight line distances

Standardized Accessibility	Population							
	Urban		Sub.		Rural		Total	
Less than 0.5	18707	6.61%	10451	32.34%	39020	25.80%	68178	14.61%
0.5 to 1	139118	49.14%	9289	28.74%	57675	38.13%	206082	44.16%
1 to 1.5	99055	34.99%	5679	17.57%	34043	22.51%	138777	29.74%
1.5 to 2	13276	4.69%	3631	11.23%	11760	7.77%	28667	6.14%
More than 2	12951	4.57%	3269	10.11%	8770	5.80%	24990	5.35%

(b) Travel times

Less than 0.5	51705	18.26%	11636	36.00%	49955	33.02%	113296	24.28%
0.5 to 1	112160	39.62%	7782	24.08%	41330	27.32%	161272	34.56%
1 to 1.5	73043	25.80%	5230	16.18%	28585	18.90%	106858	22.90%
1.5 to 2	34640	12.24%	3866	11.96%	17208	11.38%	55714	11.94%
More than 2	11559	4.08%	3805	11.77%	14190	9.38%	29554	6.33%

Table 7.9: Distribution of population counts of existing J&B accessibility to level C services. Accessibility is standardized so that the mean accessibility is one.

Although these indicators provide information on the accessibility in the existing system, it is also useful to be able to quantify these impacts in terms that are more easily interpretable by planners. One method of doing this is to examine the distribution of accessibility within the population, particularly the proportion of the population having low accessibility. Table 7.9 shows the number and percentage of women in each population group summarized by their standardized accessibility to level C services. As noted previously, the average accessibility is a constant and for level C services $\overline{A} = 0.2997$. Therefore, for ease of interpretation, accessibility has been standardized so that the average accessibility is one. Equally, for the purposed of this discussion, low

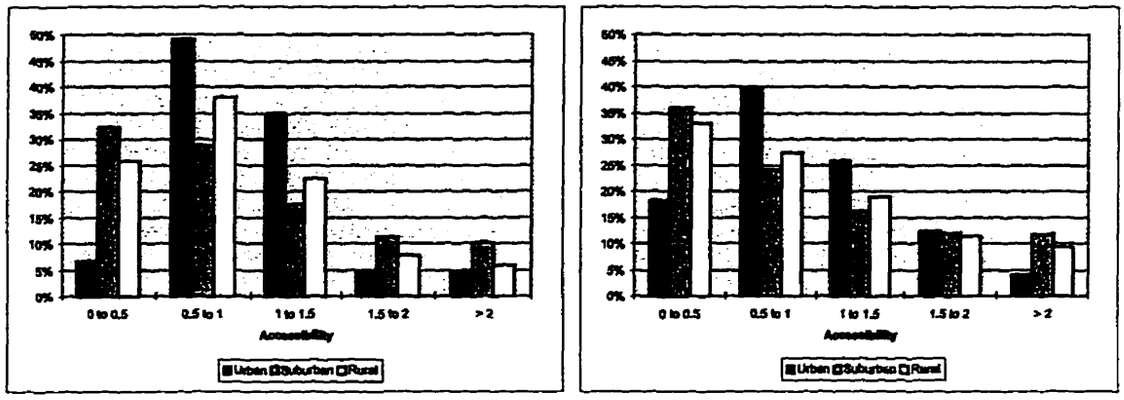


Figure 7.12: Current percentage distribution of population categorized by accessibility for *(a)* straight line distances and *(b)* estimated road network travel times.

accessibility is defined as having accessibility of less than half the average accessibility, or less than 0.5.

Table 7.9 indicates a large disparity between the population groups having low accessibility. Using straight line distances, 15% of the target population can be considered to have low accessibility. However, less than 7% of the urban population has low accessibility, while 32% of the suburban population and 26% of the rural population has low accessibility. Another interesting feature is that a much larger percentage of the population has low accessibility when accessibility is measured using travel times. In this instance, nearly a quarter of the population can be considered to have low accessibility. This is to be expected as the coefficient of variation for the travel time model is larger than with straight line distances indicating a larger variation in accessibility. Further, the disparity between population groups is also present with only 18% of the urban population having low accessibility compared with 36% of the suburban population and 33% of the rural population. Therefore, using both straight line distances and travel times, a much larger proportion of the population has low accessibility in suburban and rural areas than in urban areas.

Figure 7.13 which also illustrates these observations shows level C accessibility surfaces measured using both straight line distances and travel times with the grid cells are shaded according to standardized accessibility. In these maps, the areas considered to have low accessibility are highlighted in a pink. In contrast to the minimum distance accessibility surfaces, these surfaces show that areas with the highest accessibility are

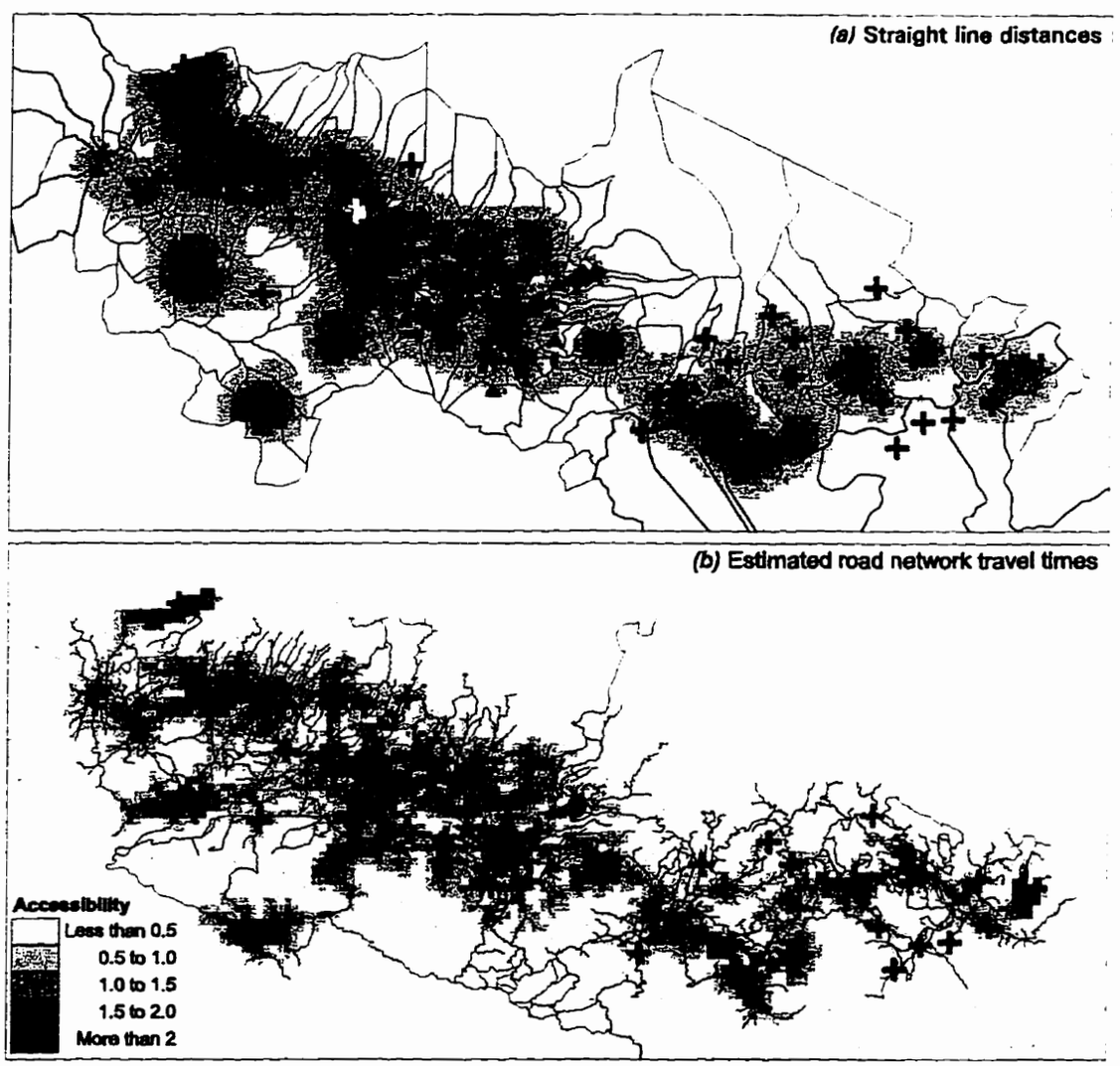


Figure 7.13: Existing standardized J&B level C accessibility *(a)* measured using straight line distances and *(b)* using estimated road network travel times.

located outside of the San José region. Though there are many large facilities in San José, there is also a very large potential demand on these facilities. Thus, though not under-supplied, this region does not exhibit very high levels of accessibility. Instead, the area with the highest accessibility is directly to the west of San José towards Alajuela and several other pockets scattered throughout the study area. The areas with the lowest accessibility are concentrated around the border of the study area and in the eastern suburbs between San José and the former capital city of Cartago.

Although the accessibility indicators and surfaces described previously provide useful visual and numerical information on the pattern of health care accessibility, an important issue is to identify strategies for improving accessibility. However, as opposed to the minimum-distance accessibility measure, J&B accessibility depends both on the locational configuration of facilities as well as the allocation of resources to these facilities. Consequently, in the previous chapter, two accessibility optimization problems were formulated, the facility location subproblem (FLS) for determining an optimal facility configuration and the resource allocation subproblem (RAS) for finding an optimal allocation of resources. Each of these optimization models is applied to the current data set. As before, they are applied using both straight line distances and road network travel times. Further, both subproblems are applied for the same two specific optimization scenarios as described in the previous section.

7.4.3 Optimizing Facility Locations

In the first instance, a three-level hierarchical FLS is applied to this data set for the two different optimization scenarios using both straight line distances and travel times. Recall that the FLS, described in Section 6.3.2, assumes that the resource allocations at each facility are fixed but that the locations of facilities can be modified so as to improve the optimization objectives. Consequently, the FLS requires specifying the size of a facility of a given type. The approach taken here was to define the facility size for a given type of facility to be equal to the average size of all existing facilities of that type. Thus, hospitals were assumed to provide 1143.9 annual hours of family planning service consultations, clinics to provide 1118.1 hours, and health centres to have 1229.6 hours. For the existing facilities in the five additional facilities scenario, the actual family planning consultation hours at each facility in 1992 were used as the facility size.

The standard FLS formulation, defined is Section 6.3.2, considers only a single level of service. In contrast to the minimum distance accessibility measure, changing the location of higher-level facilities modifies the lower-level accessibility. For example, changing the location of a hospital modifies level B and level C accessibility as well as level A accessibility. Thus, in order to analyze this three level problem, a top-down approach is used. First, the FLS is solved considering only hospitals. Next, the FLS is applied to determine the facility configuration of the clinics assuming that hospitals are fixed and located at the sites identified in the top level FLS solution. Finally, with the locations of the hospitals and clinics fixed, this process is repeated at the lowest level to find the locations of the health centres.

The FLS Interchange heuristic, described in the previous chapter, is used for solving the problems. For each facility type, the appropriate distance decay parameters are used to calculate the γ_{ij} values. Further, the resource level of a facility located at a candidate facility site, if a facility is located there, is set to be the average size of the facility type being located. For the additional facilities scenario, the Add heuristic is used to determine the initial locations of the new facilities and then the Interchange heuristic is applied to this initial solution.

Full Optimization Scenario

The full optimization scenario assumes that there are no fixed facilities and determines a new optimal facility configuration of 14 hospitals, 58 clinics, and 48 health centres of constant size, namely 1143.9, 1118.1 hours, and 1229.6 hours respectively. This approach is used for both straight line distances and road network travel times and both the efficiency ($\omega = 1$) and equity ($\omega = 0$) cases. On a 200 MHz Pentium Pro PC with 32 Megabytes of RAM and running Windows NT[9], the FLS heuristic using straight line distances required 1 hour and 6 minutes to calculate the efficiency solution and 58 minutes for the equity solution, while for road network travel times the execution times were 52 minutes and 27 minutes respectively. Although it is difficult to compare execution times for the different CPUs, the FLS model's execution time is certainly comparable to those obtained using the p-median model.

The average satisfaction and coefficient of variation for these full optimization FLS solutions are shown in Table 7.10. This table also shows the percentage improvement

[9]Approximately 40% faster than a 166 MHz Pentium.

Service Level	Average satisfaction Urban	Sub.	Rural	All	Coefficient of variation Urban	Sub.	Rural	All
(a) Straight line distances								
Efficiency solution								
Level A	-3.4225	-3.3038	-3.3180	-3.3804	0.0891	0.1413	0.1917	0.1346
	-269.02%	167.01%	117.82%	90.54%	76.16%	40.24%	53.05%	64.36%
Level B	-1.9408	-1.5668	-1.6661	-1.8258	0.2700	0.4778	0.6593	0.4482
	-9.16%	196.34%	140.58%	61.00%	49.62%	34.39%	-3.85%	23.31%
Level C	-1.3658	-1.0085	-1.1618	-1.2749	0.2440	0.5461	0.5352	0.3868
	5.37%	183.31%	121.76%	61.83%	55.79%	48.85%	24.18%	40.68%
Equity solution								
Level A	-3.4131	-3.2896	-3.3390	-3.3805	0.0853	0.1526	0.1719	0.1249
	-237.28%	180.89%	110.92%	90.45%	77.17%	35.44%	57.91%	66.92%
Level B	-1.8423	-1.5536	-1.9089	-1.8439	0.2611	0.5181	0.5090	0.3793
	48.02%	203.21%	27.05%	51.34%	51.28%	28.86%	19.83%	35.10%
Level C	-1.3208	-1.0598	-1.2816	-1.2901	0.2485	0.4818	0.5246	0.3778
	31.79%	161.56%	61.18%	53.59%	54.97%	54.87%	25.68%	42.06%
(b) Road network travel times								
Efficiency solution								
Level A	-3.4849	-3.3006	-3.2981	-3.4116	0.2039	0.2963	0.3818	0.2803
	-176.03%	169.59%	126.90%	67.43%	51.64%	2.16%	21.29%	35.78%
Level B	-1.9789	-1.8563	-1.8738	-1.9363	0.3909	0.6257	0.9137	0.5511
	18.30%	74.06%	57.14%	36.07%	38.54%	16.47%	15.38%	32.27%
Level C	-1.5019	-1.2819	-1.2983	-1.4206	0.4074	0.6950	1.0574	0.7034
	9.08%	83.45%	67.05%	33.05%	32.49%	31.78%	7.90%	17.11%
Equity solution								
Level A	-3.4596	-3.3183	-3.3470	-3.4133	0.1945	0.2911	0.3653	0.2685
	-113.99%	152.38%	109.14%	66.03%	53.87%	3.88%	24.68%	38.48%
Level B	-1.8815	-1.8917	-2.1279	-1.9621	0.4223	0.6188	0.7222	0.5511
	53.49%	65.18%	-32.91%	27.07%	33.61%	17.39%	33.11%	32.27%
Level C	-1.4048	-1.3572	-1.5944	-1.4630	0.4476	0.6609	0.7923	0.5960
	38.79%	67.28%	-37.39%	19.92%	25.82%	35.12%	31.00%	29.76%

Table 7.10: Accessibility indicators and percentage change from existing values to ideal values for the maximum efficiency and maximum equity solutions of the full FLS optimization scenario.

compared to the existing system. For the coefficient of variation, where the minimum possible value is zero, the percentage improvement, F_{CV}, is calculated as

$$F_{CV} = 1 - \frac{Z_{CV}^{NEW}}{Z_{CV}^{CUR}}$$

where Z_{CV}^{NEW} is the optimized value and Z_{CV}^{CUR} is the value in the existing system. For average satisfaction, the upper bound is $Z_{AS,k}^{MAX} = \ln\left(\sum_{\ell=1}^{k} Q_\ell / P_T\right)$ for level k services so that the percentage improvement, $F_{AS,k}$, of the average satisfaction for level k services is calculated as

$$F_{AS,k} = \frac{Z_{AS}^{NEW} - Z_{AS}^{CUR}}{\left|Z_{AS,k}^{MAX} - Z_{AS}^{CUR}\right|}.$$

Note that while the total aggregate average satisfaction cannot exceed $Z_{AS,k}^{MAX}$, a given population group's average satisfaction can, and does in several instances, exceed this value. For example, using straight line distances, the average satisfaction of the urban population to existing level A services, -3.342, was slightly greater than the upper bound of -3.372. For the full efficiency solution, the urban average satisfaction was reduced to -3.422, for a change of -269%. Therefore, the percentage improvements reported for the average satisfaction indicator must be interpreted with some caution.

As indicated in Table 7.10, the solutions to the FLS model produced large improvements in the average satisfaction and reduced the coefficient of variation relative to current potential accessibility values. As expected, the efficiency solutions increased the average satisfaction more than the equity solution. However, the equity solution reduced the coefficient of variation more. In terms of level A services, the average satisfaction improved over 90% (relative to the upper bound of 3.372) for the straight line distance solutions and by over 65% for the travel time solutions. Further, the coefficient of variation was reduced by approximately 65% (straight line distance) and 35% (travel times). In addition, the urban population experienced both the largest improvement in the coefficient of variation and the largest reduction in average satisfaction. In fact, the average satisfaction in urban areas was dramatically reduced to well below the upper bound while the average satisfaction for the suburban and rural populations increased and exceeded the upper bound for all solutions. This rather surprising result is most likely due to the effect of the potential demand on facilities located in densely populated urban areas with a correspondingly reduced resource availability. Thus, the

optimization model increased the overall aggregate satisfaction by locating facilities at sites with smaller potential demands, *i.e.*, suburban and rural areas.

A similar effect was noted in level B accessibility for the two efficiency solutions. In this case, urban areas had either a decrease or the smallest increase in average satisfaction while much larger increases occur in suburban and rural areas. However, this was not the case for the two equity solutions in which the rural population had either the smallest increase, for the straight line case, or the largest decrease, for the travel time case, in average satisfaction. Further, the reduction in the coefficients of variation for the entire target population ranged between 23% and 35%. The urban population had the largest reduction while the rural population had the smallest, except for the travel time equity solution. Compared to level A services, the increases in the efficiency and equity objectives for level B services were typically smaller.

For level C services (all facilities), the suburban population had the largest gains in average satisfaction. In fact, for the two straight line distance solutions, the average satisfaction for the suburban population was much larger than for any other population group. Overall, their average satisfaction increased by between 20% and 62% with a larger increases for the straight line distance model. These same solutions also gave a larger reduction, 40%, in the coefficient of variation as compared to the travel time models. The urban areas typically had the lowest coefficients of variation while the rural areas had the highest.

	Efficiency Solution			Equity Solution		
	Pop.	Percent	Change	Pop.	Percent	Change
Straight Line Distances						
Urban	4817	1.70%	-74.25%	5170	1.83%	-72.36%
Suburban	1266	3.92%	-87.89%	1903	5.89%	-81.79%
Rural	11625	7.69%	-70.21%	23317	15.41%	-40.24%
Total	17708	3.79%	-74.03%	30390	6.51%	-55.43%
Road Network Travel Times						
Urban	38311	13.53%	-25.90%	37691	13.31%	-27.10%
Suburban	5880	18.19%	-49.47%	6472	20.03%	-44.38%
Rural	35857	23.70%	-28.22%	52127	34.46%	4.35%
Total	80048	17.15%	-29.35%	96290	20.63%	-15.01%

Table 7.11: Total and percentage change of target population with low accessibility for full FLS optimization solutions.

Another way of assessing the FLS solutions is to examine their impact on the population with low accessibility. Using the definition of low accessibility adopted previously[10], Table 7.11 shows, for each of the four FLS solutions, the total population with low accessibility, their percentage within each population group, and the percent change in the number of people with low accessibility compared with the existing system. All of the solutions reduced the number of people with low accessibility with the efficiency solutions having the larger decrease. Further, the reduction for the straight line distance solutions was much larger than for the travel time models. This trend is a reflection of the presence of relatively isolated populations who need to travel longer distances due to poor and circuitous roads. These factors obviously do not affect the straight line distance solutions and, consequently, there were much larger reductions in the population with low accessibility. In fact, the straight line distance efficiency solution made dramatic reductions of almost 75% and lowered the total population with low accessibility from 68 178 to 17 708. Further, in suburban areas, the 32% of the population with poor accessibility in the existing system was reduced to only 4% for the same FLS solution. Although somewhat smaller, the travel time efficiency solution also reduced the population with low accessibility by over 33 000 people or by approximately 30%. Again, the suburban population experienced the largest reduction. The travel time equity solution was much less effective and only reduced the population with low accessibility by 15% and actually increased the number in rural areas.

It is also important to consider the change in system configuration and potential accessibility to family planning services from a spatial standpoint. Figure 7.14 illustrates the two surfaces showing the change in existing level C accessibility resulting from the FLS efficiency solutions for both straight line distances and travel times. The figure also shows the current and optimal facility configurations for these two cases. Overall, the two surfaces are reasonably similar, with the same regions experiencing the largest decreases and increases in accessibility. The areas showing the largest decreases are those that had very high existing accessibility.

One difficulty with the straight line solution was that it located several facilities near the edge or outside the populated area. These sites were selected because they had a low potential demand so that they increased accessibility dramatically around the edges of the grid. Further, there is a marked concentration of facilities around San José. In

[10] Accessibility of less than half the average.

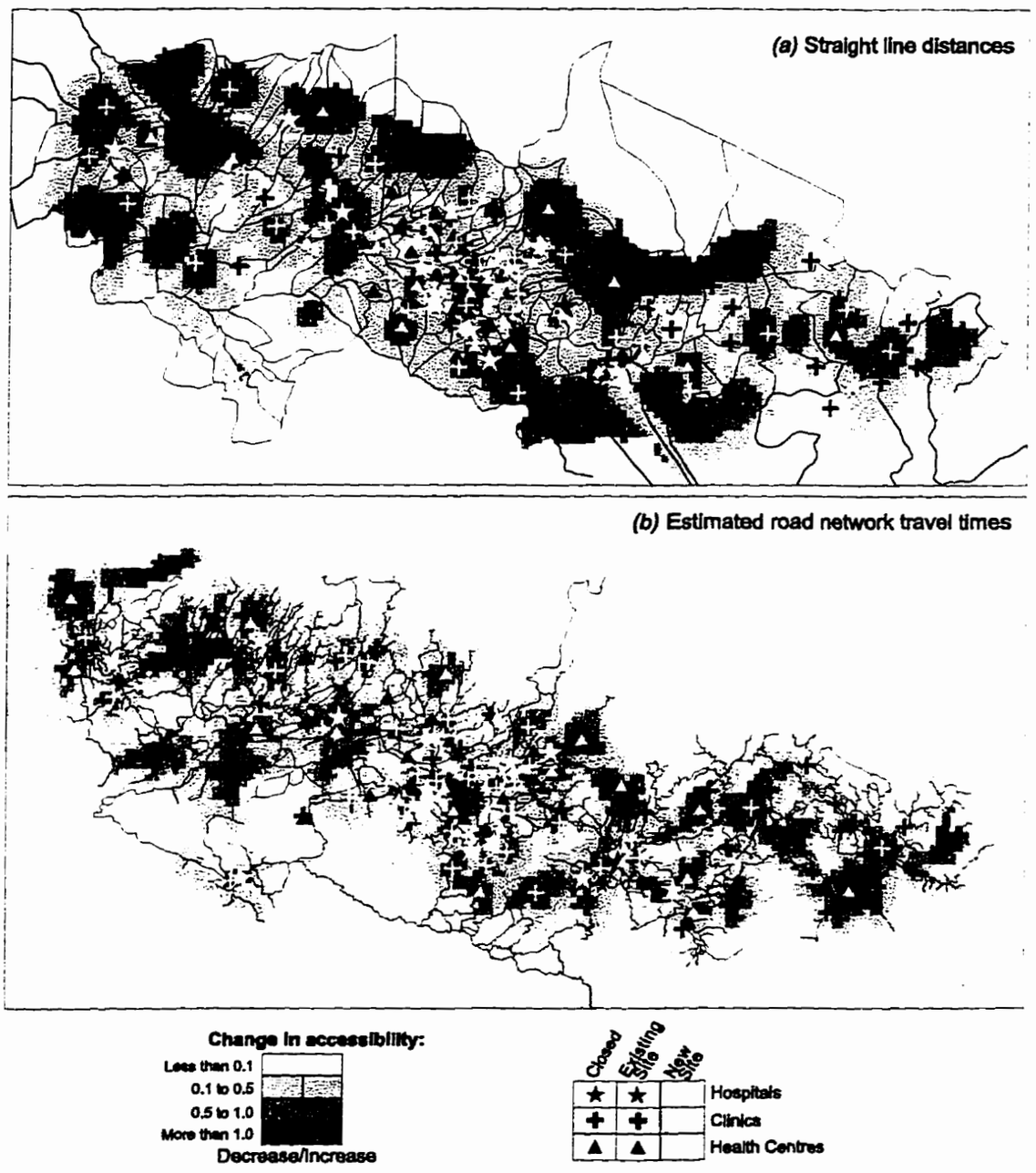


Figure 7.14: Change in J&B level C accessibility and facility locations for the full FLS optimization scenario using *(a)* measured using straight line distances and *(b)* using estimated road network travel times.

the existing system the facilities in this area are relatively large due to the concentration of population. However, for these model runs, the size of each facility was set to the average size of the existing facilities. Thus, many more smaller facilities were located in this region than in the existing system. These facilities are more dispersed than the existing configuration, with fewer located in the central core and more in suburban areas.

As indicated by this analysis, an optimal configuration of facilities can improve the average satisfaction of the target population and reduce the variation in the accessibility in comparison to the existing health care delivery system. Further, the solutions resulted in a large decreases in the number of people with low accessibility. Another approach is to apply the FLS to determine the best locations in which to open new facilities. This is achieved in the next section.

Five Additional Facilities Optimization Scenario

The second FLS optimization scenario attempts to improve the average satisfaction and reduce the coefficient of variation through the addition of five new facilities. Specifically, two new clinics and three additional health centres were located, assuming that all existing facilities remain in their current locations and are sized according to their actual family planning consultation hours in 1992. The new facilities are assumed to be the same size as the average size of the existing facilities of that type. As with the other analyses in the chapter, the model is applied using both straight line distances and road network travel times and the efficiency and equity solutions are calculated. For these model runs, the execution times, on a 166 MHz Pentium PC running Windows NT, were 528 seconds for the efficiency solution using straight lines and 451 seconds for the equity solution. The time to calculate the efficiency and equity solutions were 498 seconds and 352 seconds respectively using the network travel times. The accessibility indicators of these solutions and the percentage change from the existing system are given in Table 7.12. As previously, level A results are omitted as the configuration of hospitals is unchanged. Note that with the addition of more resources, the average accessibility of level B services increases to 0.1781, the average accessibility of level C services becomes 0.3124, and the upper bounds for the average satisfaction increase to -1.723 and -1.163 for level B and level C services respectively. These new bounds are

Service Level	Average satisfaction Urban	Sub.	Rural	All	Coefficient of variation Urban	Sub.	Rural	All
(a) Straight line distances								
Efficiency solution								
Level B	-1.8625	-1.8992	-1.9203	-1.8838	0.4956	0.6930	0.6070	0.5493
	31.32%	21.24%	19.25%	26.20%	7.52%	4.84%	4.39%	6.02%
Level C	-1.3407	-1.0991	-1.3237	-1.3185	0.5237	0.9674	0.6687	0.6133
	16.16%	123.19%	32.97%	31.08%	5.11%	9.39%	5.27%	5.95%
Equity solution								
Level B	-1.8625	-1.8992	-1.9203	-1.8838	0.4956	0.6930	0.6070	0.5493
	31.32%	21.24%	19.25%	26.20%	7.52%	4.84%	4.39%	6.02%
Level C	-1.3375	-1.1099	-1.3285	-1.3188	0.5224	0.9692	0.6699	0.6132
	17.70%	119.28%	30.99%	30.94%	5.35%	9.21%	5.11%	5.96%
(b) Road network travel times								
Efficiency solution								
Level B	-1.9879	-2.0373	-1.9776	-1.9880	0.6066	0.7112	1.0506	0.7849
	13.68%	26.83%	18.56%	16.48%	4.63%	5.05%	2.70%	3.55%
Level C	-1.4774	-1.4269	-1.4388	-1.4614	0.5779	0.9566	1.1079	0.8148
	14.71%	48.02%	15.25%	18.08%	4.23%	6.10%	3.50%	3.98%
Equity solution								
Level B	-1.9821	-2.0491	-1.9883	-1.9888	0.6045	0.7111	1.0489	0.7832
	15.59%	24.05%	15.13%	16.24%	4.95%	5.06%	2.86%	3.76%
Level C	-1.4747	-1.4407	-1.4488	-1.4640	0.5780	0.9547	1.1070	0.8142
	15.44%	45.29%	12.17%	17.37%	4.21%	6.29%	3.58%	4.04%

Table 7.12: Accessibility indicators and percentage change from existing values to ideal values for the maximum efficiency and maximum equity solutions of the add five FLS optimization scenario.

reflected in the coefficients of variation and the percentage improvement values in the table.

These results indicate that the efficiency solutions and the equity solutions are very similar. In comparison to the full optimization scenario, the reductions in the coefficient of variation in the target population are much smaller, ranging from 3.5% to approximately 6%. There are also similar modest increases for the urban, suburban, and rural populations. However, the solutions give much more dramatic improvements in the average satisfaction with increases ranging from 16% to 31%. Further, the suburban population exhibit much larger increases in average satisfaction. Recall, that the average satisfaction is based on calculating the logarithm of accessibility. Consequently, this indicator is extremely sensitive to areas with low accessibility and, in fact, would be $-\infty$ if any member of the population had zero accessibility (inaccessible). Thus, these large increases are a result of new facilities being located in areas with very low accessibility.

	Efficiency Solution			Equity Solution		
	Pop.	Percent	Change	Pop.	Percent	Change
Straight Line Distances						
Urban	13478	4.76%	-27.95%	12751	4.50%	-31.84%
Suburban	3462	10.71%	-66.87%	3604	11.15%	-65.52%
Rural	32015	21.16%	-17.95%	32354	21.39%	-17.08%
Total	48955	10.49%	-28.20%	48709	10.44%	-28.56%
Road Network Travel Times						
Urban	43246	15.28%	-16.36%	43246	15.28%	-16.36%
Suburban	8392	25.97%	-27.88%	8377	25.92%	-28.01%
Rural	46390	30.67%	-7.14%	47014	31.08%	-5.89%
Total	98028	21.00%	-13.48%	98637	21.14%	-12.94%

Table 7.13: Total and percentage change of target population with low accessibility for additional facilities FLS optimization solutions.

Further, as shown in Table 7.13, even the incremental change of adding two clinics and three health centres resulted in reasonable reductions in the population with low accessibility for all FLS solutions. For the straight line distance models, the population with low accessibility was reduced by over 19 000 people. Further, the number of people in suburban areas with poor accessibility was reduced by over 65% while in rural areas

the reduction was only about 17%. For the travel time models, while there were 15 000 fewer people with low accessibility, this was a reduction of only 13%. Nevertheless, there were 28% fewer people in suburban areas with low accessibility, 16% fewer in urban areas, and between 5% to 7% fewer in rural areas. Again, these lower numbers, compared to the straight line distance model, are likely a reflection of regions of low accessibility in isolated areas with few roads.

Figure 7.15 shows the new facility locations and the increase in accessibility for the two efficiency solutions. The straight line solution located the two new clinics near Cartago, to the East of San José, and the three new health centres in the southern and eastern fringes of the San José metropolitan area. The clinics were located near Cartago because the existing clinic in Cartago provides relatively few family planning consultations. The travel time solution showed a very similar pattern with one new clinic located in Cartago, another slightly to the West and the three health centres in a ring around San José. The two equity solutions (not shown) were practically identical to their respective efficiency solutions except for minor differences in the locations of the selected facilities sites. These solutions were also broadly similar to the solutions from the minimum distance new facilities scenario, with several of the new facilities located in the suburban areas around San José. Further, as noted previously, the population living in the suburban regions around San José is likely higher than estimated in the population grid due to the effects of in-migration where the new facilities are concentrated. Therefore, the actual improvements in accessibility are likely to be even larger than calculated here.

However, the J&B accessibility measure is also dependent on the allocation of resources to the facilities as well as their location. In the section that follows, the issue of finding an optimal allocation of resources is addressed by the resource allocation subproblem (RAS).

7.4.4 Optimizing the Allocation of Resources

For the RAS, discussed in Section 6.3.3, facility locations are assumed to be fixed and the model determines the optimal allocation of resources among the existing facilities. Although the standard RAS formulation is only for a single level of service, accessibility is defined in terms of three facility types. As before, a top-down approach is taken to solve this hierarchical problem. First, the RAS is solved to determine the optimal

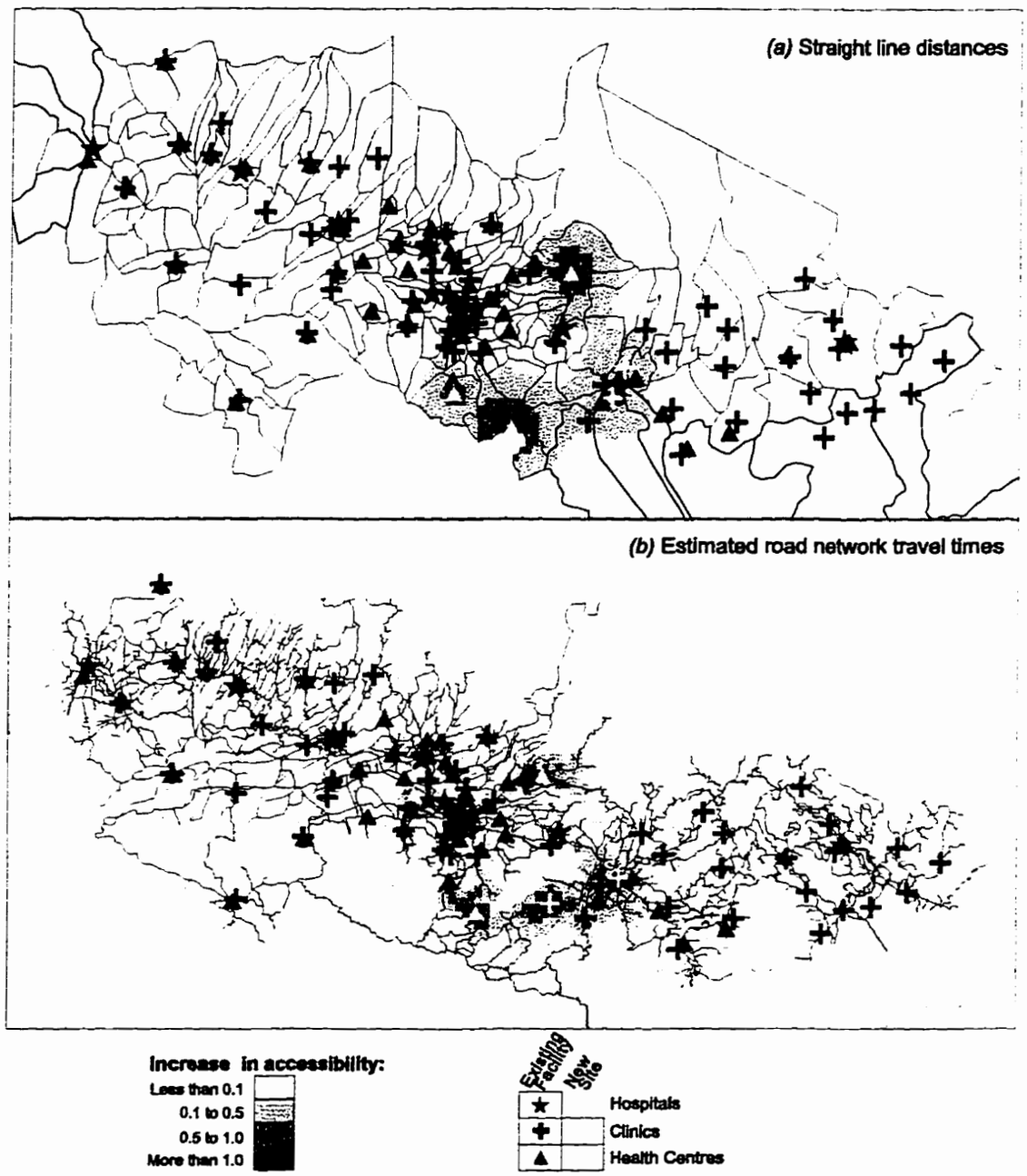


Figure 7.15: Change in level C J&B accessibility and facility locations for maximum efficiency solutions for five new facilities.

allocation of resources to the hospitals, ignoring all other facilities. Then, the sizes of the hospitals are fixed at their optimal allocations and the optimal resource allocations for the clinics are computed. Finally, with the allocation of resources to the hospitals and clinics set to their optimal values, the RAS is applied to determine the resource allocations for the health centres.

The RAS was solved using the algorithm outlined in Chapter 6. The search direction was found by solving a constrained quadratic programming (QP) problem. The QP code used was obtained from Dr. André Tits of the University of Maryland as part of the CFSQP optimization package [Lawrence *et al.*, 1996]. The line search algorithm described in the previous chapter was used and the criteria outlined by Gill *et al.* [1981, p. 100] were used to test for convergence. As opposed to the Interchange heuristic used for the FLS and the p-median problem, this algorithm solves the problem exactly so that, if the algorithm converges, the solution is optimal. One modification made to the code was that rather than recalculating the Hessian matrix after each iteration, a BFGS update[11] is used and the full Hessian is recalculated every tenth iteration[12].

Full Optimization Scenario

The full optimization scenario assumes that there are no upper or lower bounds on the allocation of resources to the facilities and that only the total level of resources for a given facility type, k, is constrained so that

$$\sum_{j \in \mathcal{F}_k} s_j = Q_k$$

where s_j is the allocation of resources to facility j and Q_k is the total level of resources for facility type k. Thus, the total resource levels were 16 015 annual hours of family planning consultation at the hospitals, 64 850 hours at the clinics, and 59 021 hours at the health centres. Further, the lower bound of resource availability at each facility was set to a small number, 1×10^{-8}, to avoid computational difficulties. The three level RAS was then solved using both straight line distances and road network travel times to find the efficiency ($\omega = 1$) and equity ($\omega = 0$) solutions. On a 166 MHz Pentium PC with 32 Megabytes of RAM running Windows NT, the problem used 137 seconds of CPU time

[11]This is a quasi-Newton method that estimates the change in the Hessian matrix from the change in gradient. See, for example, Gill *et al.*[1981, p. 116–124] for a description of these methods.

[12]This value was found, through trial and error, to give the fastest execution time on a sample problem.

to find the efficiency solution and 49 seconds to find the equity solution using straight line distances while for road network travel times, the solution times were 115 seconds and 49 seconds respectively. As expected, the equity-only RAS runs converged after one iteration for each facility type. The 49 seconds execution time mainly reflects the calculation time of the γ_{ij} values, the gradients, and the Hessian matrix. The efficiency solutions required between 6 and 23 iterations to converge.

The average satisfaction and coefficient of variation for the four optimal solutions are shown in Table 7.14 along with the percentage improvement. In general, the improvement in the indicators for the higher order services was greater than for the lower order services. Further, in comparison to the FLS solutions, these improvements were generally somewhat smaller, although still rather substantial. The reduction in the coefficient of variation ranged from 26% to 51%, again, somewhat less than with the FLS solutions. Another interesting result is the narrow range between the efficiency and equity solutions with the difference in the improvement of the indicators typically about 2% and always less than 4%. This indicates that, for the RAS, the allocation of resources to optimize either efficiency or equity was very similar. This same effect was noted in the previous chapter for the sample problem, but is less dramatic for the FLS solutions where the differences were up to 14% in certain cases.

Further, in terms of the three population groups, the RAS solutions were similar to the changes in the indicators produced by the FLS model. For example, the average satisfaction of level A services increased the most for the suburban population and decreased for the urban population. However, the improvements in the average satisfaction of level B and level C services for the suburban populations were much lower than with the FLS solutions.

In addition to improving the values of the objective functions, the new allocation of resources proposed by the RAS solutions also result in the reduction in the population with low accessibility. These figures are shown in Table 7.15. In comparison to the full FLS solutions, the RAS solutions had a somewhat smaller, but still substantial, impact on the number of people with low accessibility ranging from a reduction of 46% for the straight line distance efficiency solution to 10% for the travel time equity solution. Again, there were much larger percentage changes using straight line distances rather than travel times. For the efficiency straight line distance solution, there were over 31 000 fewer people with low accessibility than in the existing system while the corresponding reduction for the travel time efficiency solution was around 15 000

Service Level	Average satisfaction Urban	Sub.	Rural	All	Coefficient of variation Urban	Sub.	Rural	All
(a) Straight line distances								
Efficiency solution								
Level A	-3.3886	-3.3595	-3.4131	-3.3945	0.1475	0.2010	0.2383	0.1854
	-155.07%	112.38%	86.53%	74.45%	60.54%	14.95%	41.65%	50.89%
Level B	-1.8991	-1.8628	-1.7821	-1.8587	0.3729	0.5766	0.4828	0.4277
	14.99%	43.08%	86.31%	43.42%	30.42%	20.83%	23.94%	26.82%
Level C	-1.3095	-1.2455	-1.3200	-1.3085	0.3256	0.6389	0.5224	0.4255
	38.47%	82.73%	41.78%	43.56%	41.00%	40.15%	26.00%	34.75%
Equity solution								
Level A	-3.3801	-3.3596	-3.4297	-3.3948	0.1461	0.1964	0.2370	0.1839
	-126.76%	112.34%	81.08%	74.15%	60.90%	16.91%	41.96%	51.30%
Level B	-1.8797	-1.8762	-1.8281	-1.8627	0.3758	0.5676	0.4494	0.4164
	26.31%	36.16%	64.84%	41.29%	29.89%	22.07%	29.22%	28.75%
Level C	-1.2855	-1.3158	-1.3685	-1.3145	0.3412	0.5866	0.4919	0.4158
	52.58%	52.93%	17.26%	40.27%	38.17%	45.06%	30.31%	36.23%
(b) Road network travel times								
Efficiency solution								
Level A	-3.4371	-3.3979	-3.4424	-3.4361	0.2557	0.3210	0.3731	0.3032
	-59.01%	74.89%	74.48%	47.20%	39.35%	-5.99%	23.08%	30.53%
Level B	-1.9893	-2.0589	-1.8950	-1.9636	0.4899	0.6274	0.7432	0.5932
	14.53%	23.23%	49.62%	26.56%	22.98%	16.25%	31.17%	27.10%
Level C	-1.4389	-1.5997	-1.4752	-1.4618	0.5102	0.7076	0.7859	0.6267
	28.37%	15.18%	4.66%	20.29%	15.45%	30.55%	31.56%	26.14%
Equity solution								
Level A	-3.4233	-3.3921	-3.4727	-3.4372	0.2530	0.3132	0.3680	0.2992
	-25.30%	80.55%	63.48%	46.32%	40.00%	-3.42%	24.13%	31.44%
Level B	-1.9596	-2.0701	-1.9786	-1.9734	0.5075	0.6110	0.6699	0.5723
	25.28%	20.42%	20.02%	23.14%	20.20%	18.44%	37.96%	29.67%
Level C	-1.4136	-1.6270	-1.5552	-1.4743	0.5311	0.6683	0.7197	0.6082
	36.10%	9.30%	-23.58%	16.41%	11.98%	34.40%	37.32%	28.32%

Table 7.14: Accessibility indicators and percentage change from existing values to ideal values for the maximum efficiency and maximum equity solutions of the full RAS optimization scenario.

	Efficiency Solution			Equity Solution		
	Pop.	Percent	Change	Pop.	Percent	Change
Straight Line Distances						
Urban	6156	2.17%	-67.09%	8530	3.01%	-54.40%
Suburban	6168	19.08%	-40.98%	7081	21.91%	-32.25%
Rural	24217	16.01%	-37.94%	29230	19.32%	-25.09%
Total	36541	7.83%	-46.40%	44841	9.61%	-34.23%
Road Network Travel Times						
Urban	43625	15.41%	-15.63%	41429	14.63%	-19.87%
Suburban	9727	30.10%	-16.41%	10133	31.35%	-12.92%
Rural	45354	29.98%	-9.21%	49517	32.73%	-0.88%
Total	98706	21.15%	-12.88%	101079	21.66%	-10.78%

Table 7.15: Total and percentage change of target population with low accessibility for full RAS optimization solutions.

people.

One interesting trend is that the largest percentage reductions were for the urban population[13] while suburban areas experienced the largest percentage reductions with the FLS solutions. In fact, the reductions in the suburban population with low accessibility were smaller for these RAS solutions than the FLS solutions from the additional five facilities scenario. This trend is due to the fact that the RAS solutions do not change the facility configuration. Thus, areas that are distance from a facility do not experience much gain in accessibility. One of these areas includes the suburban region around San José where new facilities were located in the additional facilities FLS solutions. Thus, the RAS could not make as large of a reduction in the population with low accessibility since facilities are not located nearby while the urban areas experienced the largest reduction because these areas typically do have facilities located in close proximity. In urban areas, the problem was that the facilities had insufficient resources allocated to them relative to their potential demand. The RAS addressed this problem by allocating additional resource from areas that were relatively over-supplied to areas that were relatively under-supplied.

The spatial pattern of the change in level C accessibility as well as the percentage change in the allocation of the resources to the facilities are illustrated in Figure 7.16.

[13]Except for the travel time efficiency solution, where the suburban and urban changes were approximately the same.

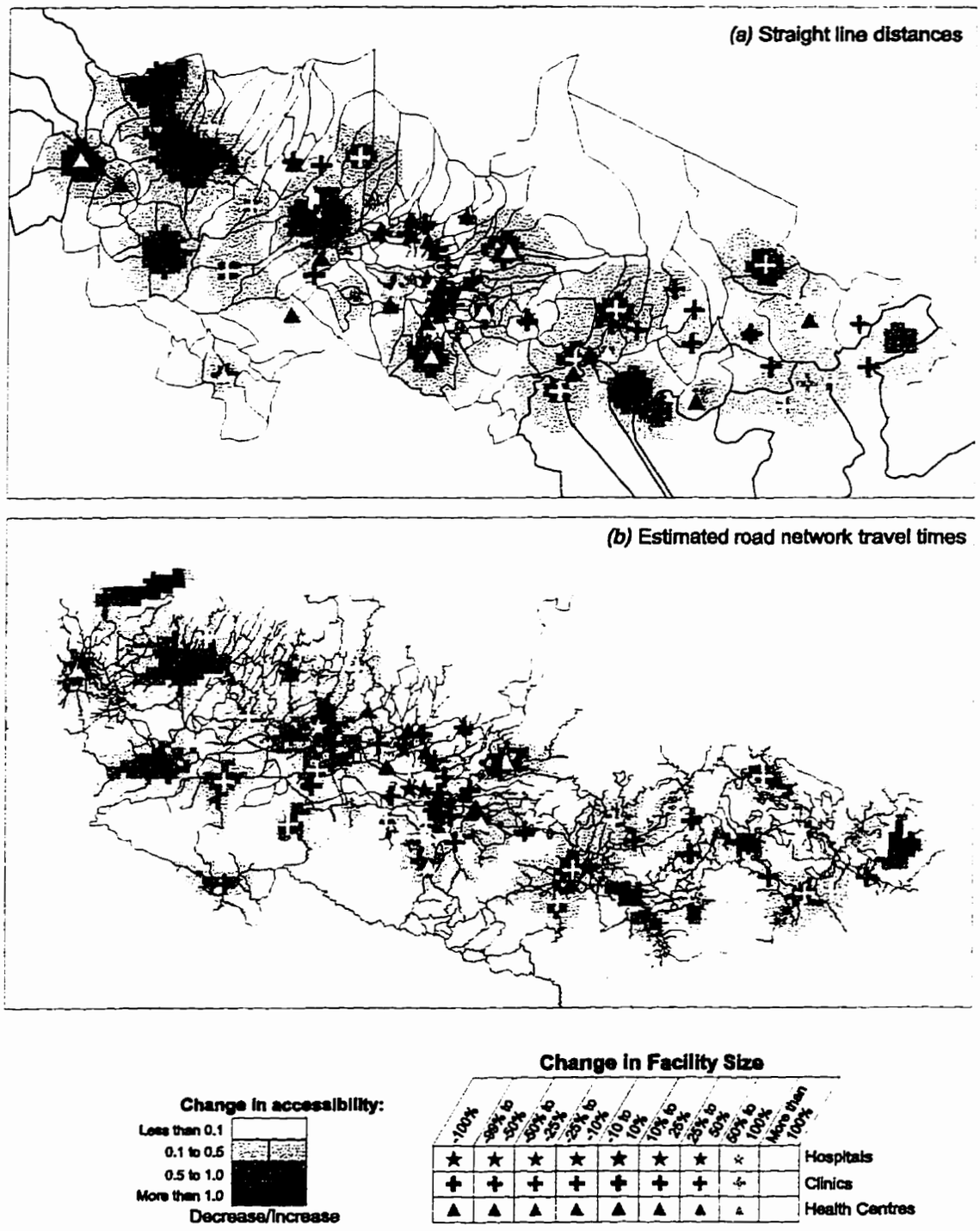


Figure 7.16: Change in J&B level C accessibility and resource allocations for the full RAS optimization scenario using *(a)* measured using straight line distances and *(b)* using estimated road network travel times.

This figure illustrates both efficiency solutions for straight line distances and road network travel times. Again, the two surfaces are reasonably similar with the largest decreases in accessibility occurring in the areas identified as having very high current accessibility. The change in resource allocations for the two solutions are also very similar. In rural areas, facilities located in areas identified as over-supplied were reduced in size while the other facilities were increased in size. Another interesting pattern, consistent with the FLS solutions, is that the facilities located in central San José were dramatically reduced in size and the resources were reallocated to facilities located in the region peripheral to San José.

Thus, even assuming that the facility locations are fixed, the solutions to the RAS model indicate that both the average satisfaction can be substantially increased, and the variation in accessibility can be reduced, by modifying the allocation of resources to facilities. Moreover, these allocations results in reductions in the population with low accessibility. However, as with the other accessibility optimization models, the RAS can examine where best to allocate additional resources so as to improve the accessibility indicators. This application of the RAS is considered in the next section.

Additional Resources Optimization Scenario

The final optimization scenario discussed in this chapter, applies the RAS model to determine the optimal allocation of additional resources to the existing facilities. Specifically, the equivalent of two average-sized clinics and three average-sized health centres were added to the existing resource levels so that the total resource levels in the optimal solutions were the same as in the FLS additional facilities scenario. The total annual hours of family planning consultations at clinics were increased by 2236.2 hours to a total of 67086.2 hours while at health centres the hours were increased by 3688.8 for a total of 62709.8 hours and the total number of hours available at hospitals remained constant. The lower bound of resource availability at each facility was set equal to the current (1992) level so that the RAS model could only increase the consultation hours at the facilities. The model was, once again, solved for the efficiency and equity solutions using both straight line distances and road network travel times. The solution times for these problems, on a 166 MHz Pentium PC with 32 Megabytes of RAM and running Windows NT, were 49 seconds for both equity solutions, 106 seconds for the travel time efficiency solution, and 129 seconds for the straight line distance efficiency

Service Level	Average satisfaction				Coefficient of variation			
	Urban	Sub.	Rural	All	Urban	Sub.	Rural	All
(a) Straight line distances								
Efficiency solution								
Level B	-1.8771	-1.8791	-1.8859	-1.8801	0.5037	0.6896	0.5941	0.5489
	24.01%	30.34%	33.53%	27.93%	6.02%	5.31%	6.42%	6.09%
Level C	-1.3250	-1.2735	-1.3394	-1.3261	0.5189	1.0064	0.6756	0.6176
	23.59%	60.26%	26.44%	27.70%	5.98%	5.74%	4.29%	5.29%
Equity solution								
Level B	-1.8776	-1.8780	-1.8856	-1.8802	0.5038	0.6895	0.5937	0.5488
	23.79%	30.84%	33.66%	27.89%	5.99%	5.33%	6.48%	6.11%
Level C	-1.3220	-1.3017	-1.3398	-1.3264	0.5179	1.0071	0.6745	0.6168
	25.00%	50.06%	26.26%	27.57%	6.16%	5.67%	4.44%	5.41%
(b) Road network travel times								
Efficiency solution								
Level B	-1.9895	-2.0770	-1.9607	-1.9862	0.6082	0.7147	1.0466	0.7841
	13.18%	17.51%	24.02%	17.05%	4.38%	4.59%	3.07%	3.64%
Level C	-1.4823	-1.5650	-1.4348	-1.4727	0.5809	0.9795	1.1156	0.8213
	13.36%	20.77%	16.48%	14.98%	3.73%	3.85%	2.84%	3.21%
Equity solution								
Level B	-1.9877	-2.0762	-1.9655	-1.9866	0.6073	0.7148	1.0462	0.7835
	13.77%	17.69%	22.49%	16.92%	4.51%	4.57%	3.11%	3.71%
Level C	-1.4800	-1.5818	-1.4375	-1.4733	0.5812	0.9764	1.1123	0.8197
	14.00%	17.47%	15.65%	14.81%	3.68%	4.16%	3.12%	3.40%

Table 7.16: Accessibility indicators and percentage change from existing values to ideal values for the maximum efficiency and maximum equity solutions of the additional resources optimization scenario.

solution. The accessibility indicators of these solutions and the percentage change from the existing system are shown in Table 7.16 (with level A results omitted) and the percentage improvement values appropriately adjusted to reflect the new upper bounds of average satisfaction.

For this optimization scenario, the efficiency and equity solutions were practically identical with very small differences in the values of the indicators between the solutions. Also, there were small reductions of less than 6.5% in the coefficients of variation. However, there were much larger gains in the average satisfaction for all cases, ranging from a minimum of 13% for the urban population in the travel time model to over 60% for the suburban population in the straight line distance model. Again, these large increases are due to the sensitivity of the average satisfaction indicator to areas of very low accessibility and the addition of resources to these areas. Further, in comparison to the FLS additional facilities scenario, the RAS had only slightly smaller improvements in the average satisfaction and the coefficient of variation.

	Efficiency Solution			Equity Solution		
	Pop.	Percent	Change	Pop.	Percent	Change
Straight Line Distances						
Urban	11331	4.00%	-39.43%	10935	3.86%	-41.55%
Suburban	7540	23.33%	-27.85%	8232	25.47%	-21.23%
Rural	33133	21.90%	-15.09%	33378	22.07%	-14.46%
Total	52004	11.14%	-23.72%	52545	11.26%	-22.93%
Road Network Travel Times						
Urban	44861	15.85%	-13.24%	44985	15.89%	-13.00%
Suburban	10384	32.13%	-10.76%	10498	32.48%	-9.78%
Rural	46200	30.54%	-7.52%	46447	30.71%	-7.02%
Total	101445	21.74%	-10.46%	101930	21.84%	-10.03%

Table 7.17: Total and percentage change of target population with low accessibility for additional resources RAS optimization solutions.

The allocation of additional resources in this optimization scenario resulted in modest reductions in the population with low accessibility. For the straight line distance models, new resource being allocated to regions that were relatively under-supplied resulted in over 15 000 fewer people having low accessibility. The number of people with poor accessibility experienced substantial decreases of 40%, of 21% to 27%, and

of 22% in urban, suburban, and rural areas respectively. For the travel time solutions, between 11 000 and 12 000 fewer people had low accessibility with the addition of new resources compared to the existing system. This resulted in a 10% decrease in the population with low accessibility. Thus, consistent with the other optimization solutions, there were smaller reductions in the population with low accessibility with the travel time solutions. Further, the suburban population experienced the smallest decrease in the number of people with low accessibility compared to the other scenarios. This is a reflection of the fact that this population lacks nearby facilities.

As can be seen in Figure 7.17, which shows the increase in resource allocations and J&B accessibility for the two efficiency solutions, the straight line distance and the travel time solutions have very similar resource allocations. The additional clinic resources were allocated around Cartago, in the eastern end of the Central Valley, and to one clinic in the southwestern section of the study area. Further, the additional resources for the health centres were three facilities located in a ring around the eastern suburbs of San José and with a moderate increase in the health centre in San Ramon at the western end of the Central Valley. Thus, although the RAS allocated some additional resources in the eastern and western end of the study area, it located most resources in the same areas that were identified by the equivalent FLS scenario.

7.5 Summary

This chapter applied two measures of geographic health care accessibility developed earlier in the thesis to evaluate family planning services using a data set consisting of women in the fertile age group in the Central Valley of Costa Rica. Further, three accessibility optimization models were applied to this data set, each for two different optimization scenarios using both straight line distance and estimated road network travel times. The health care delivery system was considered to be organized as a three-level successively-inclusive hierarchy with level A services offered at hospitals only, level B services offered at hospitals and clinics, while level C services were offered at all three types of facilities. Adjusted population counts of women in the 1992 fertile age cohort residing in urban, suburban, and rural areas were extracted from the 1984 census. These population counts were then disaggregated, using the method discussed in Chapter 4, onto a 750 metre grid which formed the base population layer for the

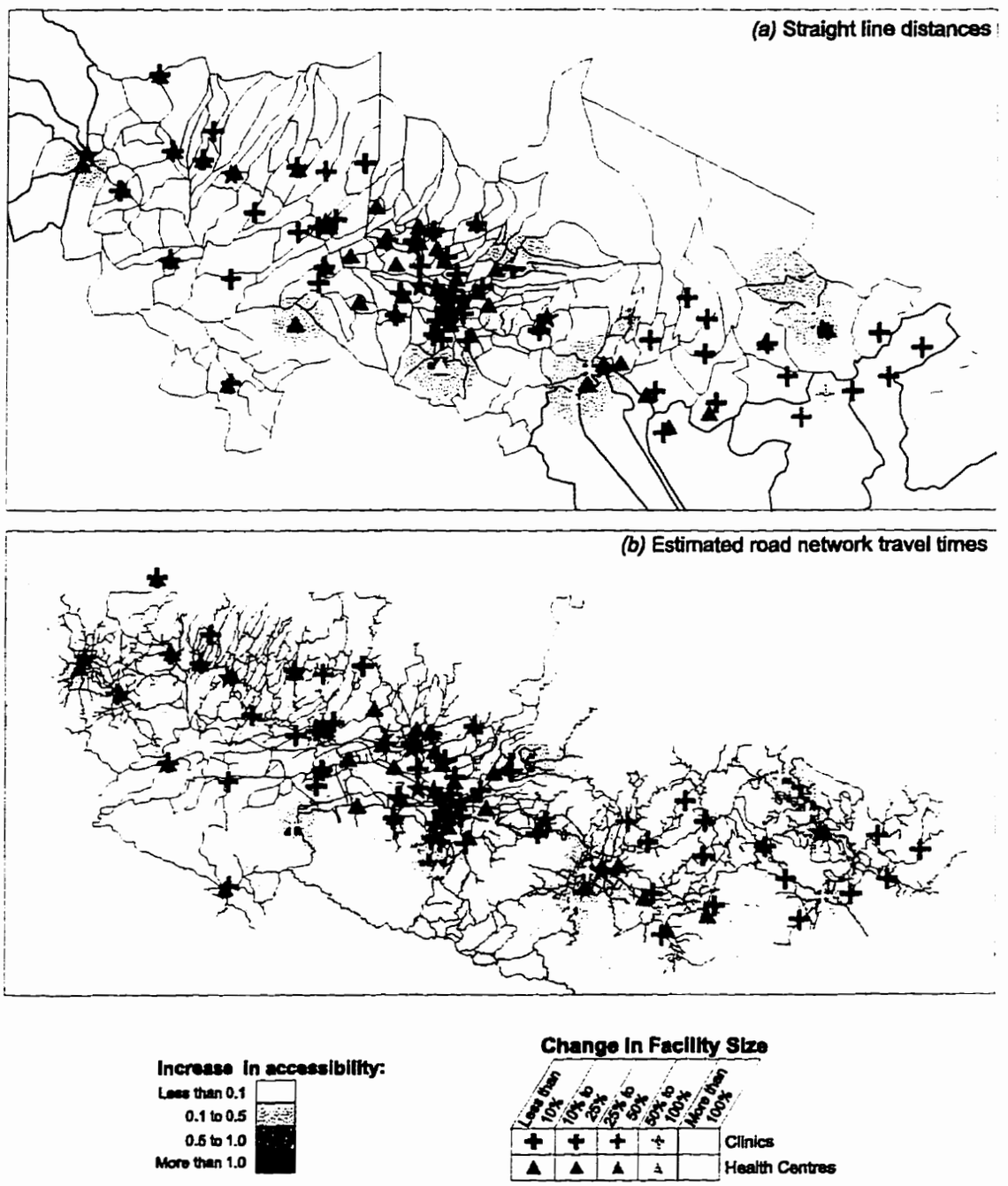


Figure 7.17: Change in level C J&B accessibility and resource allocations maximum efficiency solutions for additional resources RAS scenario.

accessibility analysis.

The minimum distance accessibility measure indicated a wide disparity between the accessibility of women living in urban areas and those of suburban and rural areas. In terms of average distance, the existing system is sub-optimally located compared to an optimal system, with the average distance to the nearest facility over 30% more. The areas with the greatest change in accessibility are located in the outer suburbs of San José and in several areas in the western Central Valley. The gains in accessibility are much more modest when five facilities are added to the existing system. Nevertheless, the largest improvements were concentrated in the suburban and rural population groups. Further, although the straight line distance and travel time solutions gave somewhat different locations for the new facilities, these facilities are located in the areas identified as under-supplied.

The J&B accessibility model was also applied to the study area using parameter estimates derived from the 1992 Costa Rican Reproductive Health Survey. The results of this analysis were less consistent than with minimum distance accessibility. In terms of the accessibility to hospitals, the urban population had the highest average access while for other services levels the measure did not find a strong urban bias although the variation in urban accessibility was lower. For level C services, the areas of highest accessibility were located outside of the urban areas in regions with large facilities but small potential demand upon them. However, a much larger percentage of the suburban and rural populations had low accessibility than did the urban population.

Further, the FLS and RAS optimization models were both applied to the study area for two different scenarios. The full optimization scenario determined an optimal facility configuration or resource allocation. The additional five facilities scenario determined the locations of five new facilities and the additional resources scenarios allocated new resources among the existing facilities. The full optimization scenario solutions indicated that large increases in average satisfaction and decreases in the coefficient of variation were possible. The gains in the accessibility indicators were more modest in the scenarios that added new facilities or resources. However, all the optimization scenarios resulted in substantial decreases in the population that had low accessibility with suburban areas experiencing the largest reductions in the FLS solutions and urban areas having the greatest decreases in the RAS solutions. Further, the solutions were broadly consistent and, as with the minimum distance accessibility analysis, identified the suburban areas of San José as under-supplied areas. The next chapter

presents a summary of the main points and contributions of this thesis, followed by a further discussion of these results, and ends with an examination of potential directions for future research.

Chapter 8

Summary and Conclusions

This chapter first reviews the contributions of the thesis in general terms and, in particular, in terms of the objectives stated in Chapter 1. This is followed by a discussion of the results of applying the accessibility measures and accessibility optimization models presented in the previous chapter. Finally, several directions for future research are outlined.

8.1 Summary

The introduction of this thesis noted the crucial importance of accessibility to primary health care services in developing countries. However, primary health care services are often inefficiently and inequitably distributed in the target population. Further, the issue of improving geographic accessibility of the target population to health care is a complex spatial problem. Thus, this thesis examined issues and models relating to the evaluation of current potential geographic accessibility and examined a facility-oriented approach to improve the spatial efficiency and equity of primary health care services using optimization models. Specifically, the objectives of this thesis were to describe a generic model of accessibility to primary health care services, examine the effects of spatial aggregation on accessibility measures, and to develop and apply accessibility optimization models. The subsequent chapters of this thesis, summarized below, fulfilled these objectives.

Chapter 2 provided a review of the issues relating to evaluating accessibility to primary health care services. It also examined two mathematical models applicable to

improving accessibility, namely, spatial interaction models and facility location models.

Chapter 3 introduced a generic model of the potential accessibility of individuals to a system of facilities providing health care services. It also provided a behavioural framework for the generic model, based on individual choice theory, and developed the minimum-distance and Joseph and Bantock [1982] accessibility measures within this framework.

However, the generic model only considered accessibility of an individual to a system of facilities. Therefore, Chapter 4 examined issues relating to the spatial aggregation of accessibility measures. It further developed some worst case error bounds on spatial aggregation error and discussed methods of disaggregating populations to a grid.

Chapter 5 discussed the use of a facility-oriented optimization approach to improve the efficiency and equity in the distribution of accessibility among the target population in a health care system and discussed some generic efficiency and equity objectives. It also introduces the generic Accessibility Optimization Problem. Two related subproblem formulations were also introduced. The Facility Location Subproblem (FLS) improves accessibility by modifying the locational configuration of the system while the Resource Allocation Subproblem (RAS) adjusts resource levels at facilities. Further, several potential optimization scenarios for each formulation were discussed.

Chapter 6 provided two specific examples of the Accessibility Optimization Problem (AOP) for the minimum distance accessibility measure and the Joseph and Bantock [1982] accessibility measure. A FLS formulation was developed for minimum distance accessibility that was equivalent to the distance-constrained p-median problem. Both FLS and RAS formulations and solution techniques were introduced for the Joseph and Bantock accessibility measure.

In Chapter 7, these accessibility measures and optimization models were applied to evaluate the current accessibility and in two specific planning scenarios that examined potential strategies to improve accessibility of family planning services in the Central Valley of Costa Rica.

8.2 Contributions

This thesis has made a contribution to current knowledge, in the areas noted above, in several ways.

1. The first contribution is of a theoretical nature and relates to the development of the generic model of potential accessibility. As indicated in Chapter 2, many different measures of potential accessibility have been proposed. However, many of these models lack a theoretical basis. The generic model provides a rigorous mathematical framework in which to consider existing measures and derive new measures and it clarifies the behavioural assumptions implicit in these measures.
2. Another useful contribution is the examination of issues relating to the errors caused by spatial aggregation of individual in accessibility measures. In particular, the worst-case error bounds for gravity model accessibility measures indicate that applying these measures to highly aggregate data can be rather problematic.
3. A further contribution is the proposed extensions to the Bracken and Martin [1989] method for disaggregating population to a grid. These particular extensions allows for the use of areal data and for the incorporation of land use classes in the estimation of a population grid.
4. Another important contribution is the generic Accessibility Optimization Problem (AOP) and its two related subproblem formulations. The AOP provides a basis for the development of suitable optimization models for taking a facility-oriented approach improving the equity and efficiency of potential accessibility.
5. A final contribution of this thesis is the formulation of the two optimization models for the Joseph and Bantock accessibility measure. These models are significant in that they allow for the examination of possible changes in either facility locations or resource allocations.

8.3 Discussion of Results

This thesis concentrated on developing mathematical models and techniques to assist health care planners and decision makers to evaluate and improve accessibility to primary health care services. The previous chapter applied these models using a sample

data set for women in the fertile age group and family planning services in the Central Valley of Costa Rica. Health care accessibility evaluation and optimization models were applied to this data set and these are discussed in turn.

8.3.1 Existing Accessibility

The evaluation of existing accessibility allows health care planners to assess the current spatial distribution of supply relative to demand and to identify areas and regions that have a deficient supply of or access to primary health care. Two different measures of potential geographic accessibility, developed within the generic model of accessibility in Chapter 3, were applied in the previous chapter, namely, the minimum distance accessibility measure, and the Joseph and Bantock [1982] (J&B) accessibility measure. These measures were applied to examine the overall accessibility of the target population to a three-level service hierarchy and the differential accessibility of women residing in urban, suburban, and rural areas. Further, these measures were applied using both straight line distances and estimated road network travel times.

In terms of minimum distance accessibility, the analysis indicated that the target population was, on average, 1.66 km and 8.3 minutes from the nearest facility. However, there was a large difference in average accessibility among the population groups with both the suburban and rural populations being over twice as far from the nearest facility as the urban population. Further, the straight line distance and estimated travel time accessibility surfaces showed that areas of high accessibility were concentrated in the major cities.

Compared with the accessibility indicators (average distance and maximum distance) for the minimum distance accessibility measure, the average satisfaction and coefficient of variation indicators, used to assess the efficiency and equity for the J&B accessibility measure, had a less straightforward interpretation. For the J&B accessibility measure, the average accessibility is constant for a given population and level of resources and, thus, is not suitable as a measure of efficiency. Instead, the average satisfaction is used to measure spatial efficiency in the distribution of J&B accessibility and is calculated as the population-weighted average of the logarithm of J&B accessibility. The coefficient of variation is a measure of equity and is calculated by dividing the standard deviation of accessibility by the mean accessibility (a constant). One method of

overcoming this difficulty is to examine the percentage of the population that has low accessibility.

The J&B accessibility measure gave similar results to the minimum distance accessibility measure with urban areas typically having the highest average satisfaction and lowest coefficient of variation. This suggests that, compared to the suburban and rural populations, the urban population had the highest level of accessibility and that this accessibility was distributed more evenly. However, the inter-group differences in accessibility were typically much lower than with minimum distance accessibility. One notable feature apparent in the J&B results was that suburban areas had both the lowest average satisfaction and highest coefficient of variation with regard to all facility (level C services) so that their accessibility was relatively poor but highly variable. One complicating factor in this analysis was that the urban population had a larger distance decay parameter, calibrated from survey data, than the suburban or rural populations indicating a stronger deterrent effect of distance. Consequently, this would have the tendency of reducing the average satisfaction of women living in urban areas and, thus, reduce the difference in accessibility between the population groups. In addition, the J&B measure considers the potential demand on a facility. This tends to reduce accessibility in densely populated urban areas while increasing accessibility in the more sparsely populated rural areas. However, there was a marked difference in the percentage of the population with accessibility of less than half of the average, defined as low accessibility, among the three population groups. Urban areas had few people with low accessibility while in rural and suburban areas the rate was between two and five times higher.

Another factor complicating the interpretations of results is the sensitivity of the average satisfaction indicator to areas of low accessibility. This can be best illustrated using a simple example with two demand nodes of equal population and an average accessibility of 1.0. If the accessibility is 0.01 in one node and 1.99 in the other then the average satisfaction is -1.96. However, the average satisfaction would increase by 58% to -0.83 if the accessibility at the nodes is 0.1 and 1.9. Thus, even a small population with low accessibility can dramatically impact on the value of the average satisfaction. In this regard, the coefficient of variation indicator is much less sensitive to small changes in accessibility. In the same situation, the coefficient of variation would decrease by only 9% from 0.99 to 0.9 still indicating a very high variability in accessibility.

The spatial pattern of current J&B accessibility indicated that the major urban areas

had average to good accessibility. Even though the facilities in these areas are relatively large so is the potential demand on them given the relatively higher urban population densities. Rather, the areas with the highest accessibility were predominantly rural. This can be explained by the fact that the population in rural areas had lower distance decay parameters than the urban population and that J&B accessibility considers both the size of a facility and its potential demand. Thus, the presence of medium-sized facilities located in rural areas that have a relatively small potential demand results in regions of high accessibility scattered throughout the rural portion of the study area.

Both the minimum distance and the J&B accessibility measures identified the same general areas as having low accessibility. These areas of low accessibility were concentrated in the area between San José and Cartago, 50 kilometres to the east, and around the boundary of the study area. Moreover, the population of these suburban areas of San José that were identified as having low accessibility is likely to be larger than the population estimates due to the effects of in-migration to the San José metropolitan area that were not considered.

Further, the accessibility surfaces were similar for accessibility measured using straight line distances and using estimated road network travel times. The straight line distance surfaces were smooth with a gradual change in accessibility between grid cells. The travel times accessibility surface exhibited a somewhat more complicated pattern with areas of higher accessibility following the road network and isolated pockets of low accessibility in grid cells distant from a road.

In summary, the two accessibility measures applied provide significant information on both the spatial distribution of accessibility and the differential accessibility between the population groups to all facilities and to subsets within the service hierarchy. The minimum distance accessibility measure found a particularly large difference in accessibility between urban areas and suburban/rural areas. This difference was less pronounced when using the J&B accessibility measure due to its consideration of facility size, the potential demand on a facility, and the different distance decay parameters for each population group. However, both measures were consistent in identifying areas of low access to health care.

8.3.2 Accessibility Optimization Models

Beyond evaluating existing accessibility, this thesis formulated a family of optimization models to assist decision makers in developing strategies to improve the efficiency and equity of potential accessibility to health care services. The previous chapter illustrated the potential use of these models by applying three specific models to the study area: the distance constrained p-median model, the J&B FLS model, and the J&B RAS model. Each of these models was applied under two specific scenarios. The first scenario generated an optimal pattern of facility locations or resource allocations for the existing system configuration. The second scenario applied the optimization models to locate additional facilities and to allocate additional resources optimally. For each scenario, two solutions were found: the efficiency solution and the equity solution.

Applying the p-median model indicated that average existing minimum distance accessibility to all facilities could be reduced by approximately 25% from the existing system. The optimal solutions also typically reduced the maximum distance of any individual to a facility. This decrease ranged from over 50% (for the urban population to hospitals) to an increase of 20% (for the efficiency solution using travel times). Further, the suburban population experienced the greatest improvement in accessibility from the optimal solutions. Much smaller reductions, ranging from 0.5% to 6%, in average distance occurred when five additional facilities were added to the existing system. Again, however, the suburban population experienced the largest reductions, up to 23%, in average distance.

The second optimization model applied to the sample data set was the J&B FLS model that changes the locational configuration of facilities so as to increase the average satisfaction (efficiency) or reduce the coefficient of variation (equity) of J&B accessibility. For the full optimization scenario, the solutions to the FLS model improved the average satisfaction by between 20% and 90% to the upper bound from the existing system accessibility while the reduction in the coefficient of variation ranged from 17% to 67% so that accessibility was more evenly distributed. Thus, an optimal facility configuration improved both of the accessibility indicators.

The suburban population experienced the largest gains in average satisfaction while the urban population experienced a relative decrease or the smallest increase with the FLS solutions. This suggests that, relative to the existing system, the optimal solutions located more facilities in suburban and rural areas and fewer in urban areas. This was

also seen the large reductions in the percentage of the population with low accessibility. Further, there was typically a large reduction in the coefficient of variation for all population groups. The urban population experienced the largest reduction and, for this group, the coefficient of variation was between 50% and 100% smaller than for the other population groups. This, again, is likely a result of there being fewer facilities located in urban areas and that these facilities are more evenly spaced. This redistribution of facilities reduces the number of women, particularly urban women, who have high accessibility as a result of residing in close proximity to several facilities. Thus, the optimal locational configuration distributes the facilities more evenly and reduces the population with very high and very low accessibility.

When five additional facilities were added, the optimal solution reduced the coefficient of variation by only modest amounts, but the increases in average satisfaction were much larger. In fact, these gains ranged between 16% and 31% and were roughly half as large as the gains in the full optimization scenario. As noted previously, the average satisfaction indicator is extremely sensitive to low accessibility values. Hence, these large gains were the result of the model locating new facilities in areas having very low initial accessibility. This was further confirmed by the large reduction in the percentage of the population with low accessibility. However, these new facilities did less to reduce the variation in accessibility as the existing configuration, with its regions of high accessibility, was unchanged.

The final accessibility optimization model was the Resource Allocation Subproblems (RAS) for the J&B accessibility measure. This model improved the efficiency and equity in the distribution of J&B accessibility by reallocating resources between facilities. In the full optimization scenario, the improvements in the average satisfaction ranged from 16% to 75% while the coefficient of variation was reduced by between 26% and 50%. Thus, the overall level of accessibility[1] increased and accessibility was more evenly distributed within the target population. This is a result of resources being moved from relatively over-supplied areas in the existing system and re-allocated to under-supplied areas.

While the suburban population experienced the largest increases in the average satisfaction for level A services (hospitals), there was no clear pattern for the other service levels with each population group receiving the largest gains in different model

[1] Accessibility measured in terms of satisfaction. The average accessibility is, of course, constant.

runs depending on whether efficiency was optimized or equity was optimized and on whether straight line distances or travel times were used. This is due to the fact that the model only reallocates resources and assumes that the facility locations are fixed. Consequently, in areas where there is no nearby facility, the accessibility cannot be improved by reallocating resources. Nevertheless, the model provided alternative resource allocations which did substantially reduce the coefficient of variation and increase the average satisfaction by moving resources from over-supplied to under-supplied areas.

The RAS was also applied to determine where additional resources should be allocated within the system so as to improve the efficiency and equity objectives. The improvements in the accessibility indicator were very similar to the additional facilities FLS scenario with reductions in the coefficient of variation ranging from 3% to 6% and increases in the average satisfaction between 15% and 28%. These increases were a result of resources being allocated to under-supplied areas concentrated in the areas between San José and Cartago.

Despite the differences in the accessibility indicators for the various optimization model solutions, overall there was a very consistent pattern for each solution. As shown in Figure 7.16, the J&B RAS solutions reduced the size of the facilities located in central San José and re-allocated the resources to suburban areas. Further, the facilities located in areas identified as having very high accessibility were reduced in size while most other facilities were increased. Both the p-median model and the J&B FLS model shown in Figures 7.10 and 7.14 respectively, removed facilities from central San José and moved them to the peripheral suburban regions surrounding the city.

Outside of the San José metropolitan area, the two facility location models adjusted the locations of facilities somewhat and tended to produce a more even spacing of facilities compared to the existing system thereby reducing the variation in accessibility and increasing the proportion of the target population with near average access. One interesting pattern in the J&B FLS solutions is the large reduction in the number of facilities in the east of the study area. For example, the efficiency J&B FLS solution for the full optimization scenario using straight line distances reduced the number of facilities east of Cartago from 24 to 9. The reduction in the number of facilities in this region is a result of the size of the facilities being located. In the existing system, the facilities located in this region are relatively small, and the FLS model located facilities of average size. Consequently, there were fewer facilities located in this region but the facilities located were larger. This, of course, is a result of the parameters used in this

optimization scenario.

Another notable tendency apparent in the results is that both facility location models tend to locate facilities in the peripheral region of the study area near the edge of the population grid cells. This pattern is because the peripheral areas were identified as having low existing accessibility. One pertinent question here is whether these areas actually have low accessibility or whether this is a result of not including facilities outside of the study area boundary in the analysis. Thus, facilities that are close to but not within the study area should perhaps have been included to reduce boundary effects. However, this inclusion introduces difficulties in the calculation of the potential demand on a facility or a candidate site for the J&B accessibility measure. This difficulty can be seen in the solution for the straight line FLS solution in Figure 7.14a where several facilities were located outside of the study area. These sites were selected because they had very little potential demand upon them. In order to get an adequate estimate of the potential demand, population information is also required for the areas surrounding the study area. Thus, for any real application of the model, care should be taken to incorporate the appropriate population and facility information for the area surrounding the study area.

Further, in the scenarios where additional resources are added to the system, each of the accessibility optimization models again produced rather consistent solutions. The facility location model solutions, shown in Figures 7.11 and 7.15, located new facilities in the suburban areas surrounding San José. Of course, there were differences between the solutions with the J&B FLS model locating the new clinics around Cartago. These differences are a function of the relatively greater sophistication of the J&B model in that it takes into account the size of the facilities. The J&B RAS solutions, shown in Figure 7.17, allocated new resources to health centres located in the suburban area of San José and to clinics located around Cartago and in the eastern area of the Central Valley. Note that, for the additional facilities or resources scenarios, the RAS solutions could adjust the sizes of many facilities while the FLS could only locate five new facilities. This explains the somewhat more dispersed pattern of accessibility change visible in the RAS solutions, as 13 facilities had additional resources allocated to them.

One feature of the J&B optimization model solutions is the similarity between the efficiency-only solutions and the equity-only solutions. This is particularly true for the RAS solutions although the efficiency and equity FLS solutions are also similar. This indicates that these two objectives are not in conflict. This feature is very similar to that

reported in Chapter 6 and shown in Figures 6.3 and 6.5. Note that a property of both objectives that they reach their optimal bound when every person in the target population has equal accessibility. This property of the objectives is due to the potential demand in the J&B measure acting as a balancing factor so that the average accessibility is a constant. The equity objective favours solutions in which the accessibility of the population is near the average accessibility while the efficiency objective favours solutions with high accessibility. However, since the average accessibility is a constant, increasing the accessibility in one area must cause a corresponding reduction in accessibility in other areas. Thus, these two objectives both favour broadly similar solutions.

Another interesting feature is the contrast between the behaviour of the J&B optimization model and the gravity-based facility location model proposed by Oppong [1992] and applied to a data set from Suhum District in Ghana. The initial version of Oppong's model located every facility in the urban area of Suhum. To overcome this difficulty, Oppong modified his model formulation so that all the population at a given location is allocated to the facility that provides the greatest benefit so that the facility location model no longer optimizes a gravity-type accessibility measure. This modified model did locate some facilities outside of the urban area but it still exhibited a very strong urban bias. On the other hand, both the FLS solutions presented in this thesis and the RAS solutions did not exhibit this urban bias and, in fact, typically reduced the average satisfaction of the urban population to less than the average satisfaction of the population living in suburban and rural areas. Although a facility located in the densely populated urban areas can increase the accessibility of a large population, it also has a high potential demand. This results in a lower increase in accessibility since the resource availability at that facility is adjusted by the potential demand. Instead, the J&B optimization models selected sites, located in suburban and rural areas, that had a lower potential demand.

The times to calculate the solutions for the two J&B optimization models were not excessive. The J&B FLS model run times were comparable to those of the p-median models and the full FLS solutions for approximately 7500 demand nodes[2] and 711 candidate facility sites took approximately one hour to compute. Birkin *et al.* [1995] reported a solution time of 150 hours for a non-linear location model on a problem with

[2]Some grid-cells were double- and triple-counted due to their being more than one population group residing in that grid-cell. Thus, each demand node represents a unique grid-cell/population group combination.

8500 demand nodes and 8000 potential sites on a Sun workstation. Oppong [1992] reported times of over 11 hours[3] to locate 29 facilities at 109 demand nodes. Although it is difficult to compare the execution times on the different computer systems, the proposed Interchange heuristic for the J&B FLS model seemed to provide a reasonably efficient solution technique. The solution times for the J&B RAS model were very reasonable with no solution requiring more than 3 minutes of calculation time.

One issue that requires further examination is the issue of the sensitivity of the models to small changes in parameter values. The J&B accessibility models require the specification of a distance decay parameter and, obviously, changes in this parameter would change the levels of existing accessibility as well as lead to different solutions for the optimization problems. A related issue is the fact that in the empirical testing reported on in the previous chapter, different distance decay parameters were used for each population group. Although these parameter values reflect the results of the calibration, they may have the tendency to replicate the existing inequity in the system. Both suburban and rural areas had much lower decay parameters than urban areas. Therefore, although an urban bias was not observed in the results, the optimization models might have allocated additional resources to the urban population due to their higher sensitivity to distance.

Several preliminary tests of the sensitivity of the model to changes in the decay parameter values were conducted to determine the strength of this effect. J&B accessibility was evaluated on the data set with the decay parameters for each population group set to be the population-weighted average of the values used in the previous chapter. The results of this analysis indicated that the same basic distribution of accessibility was evident although the accessibility in suburban and rural areas was somewhat lower. This was also reflected in the average satisfaction and coefficient of variation indicators. Urban areas had a higher average satisfaction and lower coefficient of variation than either suburban or rural areas and these differences were somewhat larger than the results reported in the previous chapter.

Further, both the FLS and RAS optimization models were tested assuming identical decay parameters for each population group. In general, the results were very similar to those reported in the previous chapter. The RAS generated a very similar pattern of resource allocation and accessibility changes with practically the same set of facilities

[3]On a 386 PC running a 33 MHz.

identified as being under-supplied, especially those in the suburban San José region. The same pattern of dramatic reductions in facilities sizes in central San José was also evident. Equally, the full FLS solution with equal decay parameters was similar to the solution reported in the previous chapter with the same areas experiencing increases in accessibility. As may be expected, the facilities were located in a slightly more dispersed pattern but there was the same trend to move facilities out of central San José and place them in the suburban area. This was also evident with the additional facilities scenario where the facilities were placed in a very similar pattern to that reported in the previous chapter; the facilities were located around Cartago and in the suburban San José region.

The models did not appear to be highly sensitive to changes in the distance decay parameters with very similar results and solutions proposed when using population group-specific decay parameter and using the same parameters for the entire population. Of course, in a planning situation, it would undoubtedly be far more appropriate to use the same decay parameters for the entire population. This would also have the advantage of reducing the influence of criteria used to define the population groups. Nevertheless, preliminary testing indicates that this does not have a large effect on the general observations of the distribution of accessibility in the study area.

Perhaps a more critical issue is the reliability of the population estimates. The model located facilities and allocated resources based on the estimated population distribution. As noted previously, there is likely to be an under-estimation of the target population in the suburban region of San José due to the effects on in-migration. Further, as discussed previously, the average satisfaction indicator is very sensitive to areas of very low accessibility. If the population estimates in these areas are either smaller or larger than in reality, this could have a strong effect on the solutions. Therefore, a critical issue for the successful application of these models in a planning situation is to develop improved methods of estimating the population distribution and the availability of reliable population counts.

8.4 Directions for Future Research

This thesis has outlined several useful mathematical models for evaluating and improving potential accessibility to primary health care services. However, there are many potential directions available for future research in this area. This section outlines some

of these areas which require further research that can extend the models and methods previously discussed.

1. Although the use of a population grid reduced the spatial aggregation error in evaluating accessibility, there were undoubtedly errors in the grid-cell population estimates. However, without an accurate estimate of the spatial distribution of demand, the results of evaluating accessibility and the solutions to the accessibility optimization models must be treated cautiously. If an area is represented as being populated but really is unpopulated, or vice versa, then the accessibility optimization model would locate facilities or allocate resources inappropriately. This is further amplified in the J&B accessibility optimization models which are very sensitive to areas of low accessibility. This underlines the importance of obtaining an accurate representation of the population.

   Ideally, this would require more disaggregate population data, such as by *segmentos* in Costa Rica. These data may be difficult to obtain in many developing countries. However, there are several areas of research that may yield improved population estimates. One such area would be to investigate the use of road network distances, or travel times, to calculate the weighting of grid cells in the Bracken and Martin [1989] method. Since settlement patterns typically follow the road network, this may produce more realistic population estimates.

   Further, the use of remotely sensed imagery in order to obtain land use classes or a preliminary population estimate for the grid cells should also be investigated. For example, Iisaka and Hegedus [1982] provide a method for estimating population using Landsat imagery. These initial estimates could be used as the population propensity of a grid cell and can be used to provide additional information to the grid-cell population disaggregation procedure.

2. For the J&B accessibility measure, two accessibility indicators were used to report the results: the average satisfaction, and the coefficient of variation. Although the coefficient of variation seemed to present a relatively clear indicator of the equity in the system, the average satisfaction was somewhat more difficult to interpret. Further research is required in order to develop an efficiency indicator for the J&B accessibility measure that can be more easily interpreted. One possible means of overcoming this would be to examine a standardized difference between the

average satisfaction and its upper bound so that this indicator, like the coefficient of variation, is comparable between different study areas.

3. Two specific subproblems, the Facility Location Subproblem (FLS), and the Resource Allocation Subproblem (RAS), were formulated for the J&B accessibility measure. However, one area that merits further investigation is developing appropriate solution techniques that can simultaneously optimize both the locations of facilities and the allocation of resources. One possible technique would be to use a two-phase algorithm that alternates between re-allocating resources and changing the location configuration of the facilities.

   In addition, the two objective functions used did not distinguish between population groups and optimized the efficiency and equity of the entire target population. It is possible to develop alternative objective function formulations that would not only optimize the objectives for the entire population but could also reduce the difference in objective values between the population groups. For example, one potential objective function formulation could attempt to reduce the difference in efficiency between population groups as well as increasing the overall efficiency.

4. Another avenue to explore with the J&B accessibility measure would be to examine an accessibility measure based on a linear, rather than an exponential, transformation from satisfaction to accessibility and develop the corresponding accessibility optimization objectives and models. Using a linear transformation, accessibility would be expressed as

$$A_i = \ln\left[H_i + \sum_j S_j / C_j \exp\left(-\beta D_{ij}\right)\right].$$

   It would be interesting to compare the results of applying this model to the standard J&B accessibility measure.

5. A related area for further research is to consider other alternative accessibility measures. The J&B accessibility measure incorporates the congestion of the facility by dividing the size by the potential demand. This led to high accessibility in rural areas near facilities with a small potential demand which may not be an accurate assessment of accessibility in this situation. Simply dividing by the potential demand may over-estimate the availability of resource at fa-

cilities with a small potential demand. Further research is required to develop formulations that incorporate congestion in accessibility measure. For example, Leonardi [1980a; 1980b] proposed that the congestion of a facility be related to the demand on a facility that is greater than a pre-defined capacity.

6. As outlined in Chapter 3, the random utility framework can incorporate clustered alternatives. Although this was not further explored in this thesis, this certainly can incorporated within accessibility measures. It would be interesting to calibrate a nested multinomial logit model to evaluate accessibility, assuming that the attractiveness of alternatives were correlated.

7. The accessibility measures applied in this thesis use very few facility dependent factors and treat service availability in a rather simplistic manner. Where appropriate data are available, accessibility measures can be developed that incorporate in a more sophisticated manner the various resources and services available at a facility. Further, it is possible to develop the appropriate optimization models from these accessibility measures.

Bibliography

[Aday and Andersen, 1974] L.A. Aday and R. Andersen. A framework for the study of access to medical care. *Health Services Research*, 9:208–220, 1974.

[Akhtar and Izhar, 1986] R. Akhtar and N. Izhar. Inequalities in the distrubution of health care in India. In R. Akhtar and A.T.A. Learmoth, editors, *Geographical Aspects of Health and Disease in India*, pages 437–460. Concept, New Delhi, 1986.

[Akin *et al.*, 1985] J.S. Akin, C.C. Griffin, D.K. Guilkey, and B.M. Popkin. *The Demand for Primary Health Services in the Third World*. Rowman & Allanheld, Totowa, N.J., 1985.

[Alonso, 1978] W. Alonso. A theory of movement. In N.M. Hansen, editor, *Human Settlement Systems: International Perspectives on Structure, Change, and Public Policy*, pages 197–211. Ballinger, Cambridge, Mass., 1978.

[Annis, 1981] S. Annis. Physical access and utilization of health services in rural Guatemala. *Social Science and Medicine*, 15D:313–523, 1981.

[Ayeni *et al.*, 1987] B. Ayeni, G. Rushton, and M.L. McNulty. Improving the geographical accessibility of health care in rural areas: A Nigerian case study. *Social Science & Medicine*, 25:1083–1094, 1987.

[Bailey and Phillips, 1990] W. Bailey and D.R. Phillips. Spatial patterns of use of health services in the Kingston Metropolitan Area, Jamaica. *Social Science and Medicine*, 30:1–12, 1990.

[Balas and Ho, 1980] E. Balas and A. Ho. Set covering algorithms using cutting planes, heuristics and subgradient optimisation: A computational study. *Mathematical Programming Study*, 12:37–60, 1980.

[Beaumont, 1980] J.R. Beaumont. Spatial interaction models and the location-allocation problem. *Journal of Regional Science*, 20:37–50, 1980.

[Beaumont, 1987] J.R. Beaumont. Location-allocation models and central place theory. In A. Ghosh and G. Rushton, editors, *Spatial Analysis and Location-Allocation Models*, pages 21–54. Van Nostrand Reinhold, New York, 1987.

[Ben-Akiva and Lerman, 1985] M. Ben-Akiva and S.R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, Massachusetts, 1985.

[Bennett *et al.*, 1982] V.L. Bennett, D.J. Eaton, and R.L. Church. Selection sites for rural health worker. *Social Science & Medicine*, 16:63–72, 1982.

[Birkin *et al.*, 1995] M. Birkin, M. Clarke, and F. George. The use of parallel computers to solve nonlinear spatial optimisation problems: An application to network planning. *Environment and Planning A*, 27:1049–1068, 1995.

[Bracken and Martin, 1989] I. Bracken and D. Martin. The generation of spatial population distributions from census centroid data. *Environment and Planning A*, 21:537–543, 1989.

[Bracken and Martin, 1995] I. Bracken and D. Martin. Linkage of the 1981 and 1991 UK censuses using surface modelling concepts. *Environment and Planning A*, 27:379–390, 1995.

[Bracken, 1993] I. Bracken. An extensive surface model database for population-related information: concept and application. *Environment and Planning B*, 20:13–27, 1993.

[Bracken, 1995] I. Bracken. The generation of spatial population distributions from census centroid data. In A.S. Fotheringham and P. Rogerson, editors, *Spatial Analysis and GIS*, pages 247–259. Taylor & Francis, 1995.

[Bradshaw, 1972] J. Bradshaw. A taxonomy of social need. In G. McLahan, editor, *Problems and Progress in Medical Care*. Oxford University Press, Oxford, 1972.

[Cadwallader, 1975] M. Cadwallader. A behavioral model of consumer spatial decision making. *Economic Geography*, 51:339–349, 1975.

[Casillas, 1987] P.A. Casillas. Data aggregation and the p-median problem in continuous space. In A. Ghosh and G. Rushton, editors, *Spatial Analysis and Location-Allocation Models*, pages 327–344. Van Nostrand Reinhold, New York, 1987.

[CCSS, 1991] *Anuario Estadístico 1991*. Caja Costarricense de Seguro Social, San José, Costa RIca, 1991.

[CCSS, 1994] *Fecundidad y Formación de la Familia: Encuesta Nacional de Salud Reproductiva*. Caja Costarricense de Seguro Social, San José, Costa RIca, 1994.

[CELADE, 1992] Latin America: Notes on population, environment and development (iesa/p/ac.34/inf.6). Paper presented at the United Nations Working Group for Population, Environment, and Development meeting, New York, 1992.

[Chakiel and Martinez, 1992] Chakiel and Martinez. Transición demografica en America Latina y el Caribe desde 1950. Paper presented at the Fourth Conference on Demographic Transition, 1992. Mexico City, March 23-26.

[Chen *et al.*, 1983] C.H.C. Chen, R. Santiso, and L. Morris. Impact of accessibility of contraceptives on contraceptive prevalence in Guatemala. *Studies in Family Planning*, 14:275–283, 1983.

[Chernichovsky and Meesok, 1986] D. Chernichovsky and O.A. Meesok. Utilization of health services in Indonesia. *Social Science and Medicine*, 23:611–620, 1986.

[Choi and Chaudry, 1993] I.C. Choi and S.S. Chaudry. The p-median problem with maximum distance constraints: A direct approach. *Location Science*, 1:235–243, 1993.

[Choukron, 1975] J.-M. Choukron. Development of gravity-type trip distribution models. *Regional Science and Urban Economic*, 5:177–202, 1975.

[Church *et al.*, 1993] R. Church, J. Current, and H. Eiselt. Editorial. *Location Science*, 1:1–3, 1993.

[Cornelius and Novak, 1983] R.M. Cornelius and J.A. Novak. Contraceptive availibility and use in five developing countries. *Studies in Family Planning*, 14:302–317, 1983.

[Current and Schilling, 1989] J.R. Current and D.A. Schilling. Analysis of errors due to demand data aggregation in the set covering and maximal covering location problems. *Geographical Analysis*, 21:116–126, 1989.

[Daganzo, 1979] C. Daganzo. *Multinomial Probit: The Theory and Its Application to Demand Forecasting*. Academic Press, New York, 1979.

[Dear, 1974] M.J. Dear. A paradigm for public facility location. *Antipode*, 6:46–50, 1974.

[Densham and Rushton, 1992] P.J. Densham and G. Rushton. Strategies for solving large location-allocation problems by heuristic methods. *Environment and Planning A*, 24:289–304, 1992.

[Dirección General de Estadística y Censo, 1992] Dirección General de Estadística y Censo. Costa Rica cálculo de población: Enero 1992. San José, Costa Rica, 1992.

[Dökmeci, 1979] V.F. Dökmeci. A multiobjective model for regional planning of health facilities. *Environment and Planning A*, 11:517–525, 1979.

[Domencich and McFadden, 1975] T.A. Domencich and D. McFadden. *Urban Travel Demand: A Behavioral Analysis*. North-Holland, Amsterdam, 1975.

[Donabedian, 1973] A. Donabedian. *Aspects of Medical Care Adminstration*. Harvard University Press, Cambridge, MA, 1973.

[Easterlin *et al.*, 1988] R.A. Easterlin, K. Wongboonsin, and M.A. Ahmed. The demand for family planning: a new approach. *Studies in Family Planning*, 19:257–269, 1988.

[Eaton *et al.*, 1981] D.J. Eaton, R.L. Church, V.L. Bennett, B.L. Hamon, and L.G.V. Lopez. On deployment of health resource in rural Valle del Cauca, Colombia. *TIMS Studies in the Management*, 17:331–359, 1981.

[Eaton *et al.*, 1986] D.J. Eaton, H.M. Sánchez, R.R. Lantigua, and J. Morgan. Determining ambulance deployment in Santo Domingo, Dominican Republic. *Journal of the Operations Research Society*, 37:113–126, 1986.

[ECLAC, 1992] *Social Equity and Changing Production Processes: An Integrated Approach.* Number LC/G.1701 (SES.24/3). Economic Commision for Latin America and the Caribbean (ECLAC), Santiago, Chile, 1992.

[ECLAC, 1993] *Population, Social Equity, and Changing Production Patterns.* Number LC/DEM/G.131. Economic Commision for Latin America and the Caribbean (ECLAC), Santiago, Chile, 1993.

[Entwisle *et al.*, 1995] B. Entwisle, R.R Rindfuss, S.J. Walsh, T. Evans, and S.R. Curran. Geographical information systems, spatial network analysis. and family planning program evaluation. Paper presented at the Annual Meeting of the American Population Association, April 1995.

[Ewing, 1976] G.O. Ewing. Environmental and spatial preferences of interstate migrants in the United States. In R.G. Golledge and G. Rushton, editors, *Spatial Choice and Spatial Behavior*, pages 249–269. Ohio State University Press: Columbus, 1976.

[Feldman *et al.*, 1966] E. Feldman, F.A. Lehrer, and T.L. Ray. Warehouse location under continuous economies of scale. *Management Science*, 12:670–684, 1966.

[Fendall, 1981] N.R.E. Fendall. Primary health care: Issues and constraints. *Third World Planning Review*, 3:387–401, 1981.

[Fielder, 1981] J.L.A. Fielder. A review of the literature on access and utilization with special emphasis on rural primary care. *Social Science and Medicine*, 15C:129–142, 1981.

[Fik and Mulligan, 1990] T.J. Fik and G.F. Mulligan. Spatial flows and competing central plances: Towards a general theory of hierarchical interaction. *Environment and Planning A*, 22:527–549, 1990.

[Fishburn, 1970] P.C. Fishburn. *Utility Theory for Decision Making.* John Wiley & Sons, New York, 1970.

[Fisher and Rushton, 1979] H.B. Fisher and G. Rushton. Spatial efficiency of service locations and the regional development problem. *Papers of the Regional Science Association*, 42:83–97, 1979.

[Fodor and Roubens, 1994] J. Fodor and M. Roubens. *Fuzzy Preference Modelling and Multicriteria Decision Support.* Kluwer, Dordrecht, 1994.

[Fotheringham and O'Kelly, 1989] A.S. Fotheringham and M.E. O'Kelly. *Spatial Interaction Models: Formulations and Applications.* Kluwer, Dordrecht, 1989.

[Fotheringham *et al.*, 1995] A.S. Fotheringham, P.J. Densham, and A. Curtis. The zone definition problem in location-allocation modeling. *Geographical Analysis*, 27:60–77, 1995.

[Fotheringham, 1981] A.S. Fotheringham. Spatial structure and distance-decay parameters. *Annals of the Association of American Geographers*, 71:425–436, 1981.

[Fotheringham, 1983] A.S. Fotheringham. A new set of spatial interaction models: the theory of competing destinations. *Environment and Planning A*, 15:15–36, 1983.

[Francis and Lowe, 1992] R.L Francis and T.J. Lowe. On worst-case aggregation analysis for network location problems. *Annals of Operations Research*, 40:229–246, 1992.

[Geertman and van Eck, 1995] S.C.M. Geertman and J.R.R. van Eck. GIS and models of accessibility potential: An application in planning. *International Journal of Geographical Information Systems*, 9:67–90, 1995.

[Gill *et al.*, 1981] P.E. Gill, W. Murray, and M.H. Wright. *Practical Optimization.* Academic Press, London, 1981.

[Girt, 1973] J.L. Girt. Distance to general medical practice and its effect on revealed ill-health in a rural environment. *Canadian Geographer*, 17:154–166, 1973.

[Gold, 1989] C.M. Gold. Surface interpolation, spacial adjacency and GIS. In J. Raper, editor, *Three Dimensional Applications in Geographical Information Systems*, pages 21–35. Taylor and Francis, London, 1989.

[Goodchild, 1979] M.F. Goodchild. The aggregation problem in location-allocation. *Geographical Analysis*, 11:240–255, 1979.

[Gore, 1991a] C.G. Gore. The spatial separatist theme and the problem of representation in location-allocation models. *Environment and Planning A*, 23:939–953, 1991.

[Gore, 1991b] C.G. Gore. Location theory and service development planning: Which way now? *Environment and Planning A*, 23:1095–1109, 1991.

[Guzman, 1992] J.M. Guzman. *Crisis, Adjustment, and Fertility during Latin America's lost decade: Facts and speculations.* Latin American Demographic Center (CELADE), Santiago, Chile, 1992.

[Habib and Vaughn, 1986] O.S. Habib and J.P. Vaughn. The determinants of health services utilization in southern Iraq: A household interview survey. *International Journal of Epidemology*, 15:395–403, 1986.

[Hall, 1988] G.B. Hall. Monitoring and predicting community mental health utilization in Auckland, New Zealand. *Social Science and Medicine*, 26:55–70, 1988.

[Halleford and Jörnsten, 1985] Å. Halleford and K. Jörnsten. A note on relaxed gravity models. *Environment and Planning A*, 17:597–603, 1985.

[Hansen *et al.*, 1983] P. Hansen, D. Peeters, and J.-F. Thisse. Public facility location models: A selective survey. In J.-F. Thisse and H.G. Zoller, editors, *Locational Analysis of Public Facility*, pages 223–262. North-Holland, Amsterdam, 1983.

[Hansen *et al.*, 1987] P. Hansen, M. Labbé, D. Peeters, and J.-F. Thisse. Facility location analysis. In Richard Arnott, editor, *Systems of Cities and Facility Location*, Fundamental of Pure and Applied Econmics, pages 1–70. Harwood Academic Publishers, Chur, Switzerland, 1987.

[Hansen, 1959] W.G. Hansen. How accessibility shapes land use. *Journal of the American Institute of Planners*, 25:73–76, 1959.

[Hart *et al.*, 1990] R.H. Hart, M.A. Beasely, and E. Tarimo. *Integration Maternal and Child Health Services with Primary Health Care.* World Health Organization, Geneva, 1990.

[Haynes and Bentham, 1982] R.M. Haynes and C.G. Bentham. The effects of accessibility on general practioner consultations, out-patient attendances and in-patient admissions in Norfolk, England. *Social Science and Medicine*, 16:561–569, 1982.

[Hellen, 1986] J.A. Hellen. Medical geography and the third world. In M. Pacione, editor, *Medical Geography: Progress and Prospect*. Croom Helm, London, 1986.

[Hillsman and Rhoda, 1978] E.L. Hillsman and R. Rhoda. Errors in measuring distance from populations to service centers. *Annals of Regional Science*, 12(3):74–88, 1978.

[Hillsman, 1984] E.L. Hillsman. The p-median structure as a unified linear model for location-allocation analysis. *Environment and Planning A*, 16:305–318, 1984.

[Hodgart, 1978] R.L. Hodgart. Optimizing access to public services: A review of problems, models and methods of locating central facilities. *Progress in Human Geography*, 2:17–48, 1978.

[Hodgson and Neuman, 1993] M.J. Hodgson and S. Neuman. A GIS approach to eliminating source C aggregation error in p-median models. *Location Science*, 1:155–170, 1993.

[Hodgson, 1978] M.J. Hodgson. Towards more realistic allocation in location-allocation models: An interaction approach. *Environment and Planning A*, 10:1273–1285, 1978.

[Hodgson, 1984] M.J. Hodgson. Alternative approaches to hierarchical location-allocation systems. *Geographical Analysis*, 16:275–281, 1984.

[Hodgson, 1988] M.J. Hodgson. An hierarchical location-alloction models for primary health care delivery in a developing area. *Social Science and Medicine*, 26:153–161, 1988.

[Iisaka and Hegedus, 1982] J. Iisaka and E. Hegedus. Population estimation from Landsat imagery. *Remote Sensing of Environment*, 12:259–272, 1982.

[Johnson and Kotz, 1970] N.L. Johnson and S. Kotz. *Continuous Univariate Distributions*, volume 1. Houghton Mifflin, Boston, 1970.

[Johnson and Papdimitrou, 1985] D.S. Johnson and C.H. Papdimitrou. Computation complexity. In E.L Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys, editors, *The Traveling Salesman Problem*, pages 37–86. John Wiley and Sons, 1985.

[Joseph and Bantock, 1982] A.E. Joseph and P.R. Bantock. Measuring potential physical accessibility to general practitioners in rural areas: A method and case study. *Social Science & Medicine*, 16:85–90, 1982.

[Joseph and Phillips, 1984] A.E. Joseph and D.R. Phillips. *Accessibility and Utilization: Geographical Perspectives on Health Care Delivery*. Harper & Row, New York, 1984.

[Joseph, 1982] A.E. Joseph. On the interpretation of the coefficient of localization. *Professional Geographer*, 34:443–446, 1982.

[Kanaroglou and Hall, 1989] P. Kanaroglou and B. Hall. A framework for the analysis of psychiatric health facility utilization. *Geographia Medica*, 19:115–140, 1989.

[Kariv and Hakimi, 1979a] O. Kariv and S.L. Hakimi. An algorithm approach to network location problems I, the *p*-centers. *SIAM Journal on Applied Mathematics*, 37:513–538, 1979.

[Kariv and Hakimi, 1979b] O. Kariv and S.L. Hakimi. An algorithm approach to network location problems II, the *p*-medians. *SIAM Journal on Applied Mathematics*, 37:539–560, 1979.

[Khan and Bhardwaj, 1994] A.A. Khan and S.M. Bhardwaj. Access to health care: A conceptual framework and its relevance to health care planning. *Evaluation & The Health Professions*, 17:60–76, 1994.

[Khan, 1985] A.A. Khan. Analyzing spatial disparities in access to health care: A methodology with application to Bangladesh. *GeoJournal*, 10:91–107, 1985.

[Khan, 1992] A.A. Khan. An integrated approach to measuring potential spatial access to health care services. *Socio-Economic Planning Sciences*, 26:275–287, 1992.

[Khumawala, 1973] B.M. Khumawala. An efficient algorithm for the *p*-median problem with maximum distance constraints. *Geographical Analysis*, 5:309–321, 1973.

[Kleczkowski and Pubouleau, 1976] B.M. Kleczkowski and R. Pubouleau, editors. *Approaches to Planning and Design of Health Care Facilities in Developing Areas*. World Health Organization, Geneva, 1976. 5 volumes.

[Knox, 1978] P.L. Knox. The intraurban ecology of primary medical care: patterns of accessibility and their policy implications. *Environment and Planning A*, 10:415–435, 1978.

[Knox, 1979] P.L. Knox. Medical deprivation, area deprivation and public policy. *Social Science and Medicine*, 13D:111–121, 1979.

[Kuehn and Hamburger, 1963] A.A. Kuehn and M.J. Hamburger. A heuristic program for locating warehouses. *Management Science*, 9:643–666, 1963.

[Lam, 1983] N.S. Lam. Spatial interpolation methods: a review. *American Cartographer*, 10:129–149, 1983.

[Lawrence *et al.*, 1996] C.T. Lawrence, J.L. Zhou, and A.L. Tits. User's guide for CFSQP version 2.4: A C code for solving (large scale) constrained nonlinear (minimax) optimization problems. Technical Report TR-94-16r1, Institute for Systems Research, University of Maryland, College Park, MD, 1996.

[Leonardi, 1978] G. Leonardi. Optimum facility location by accessibility maximizing. *Environment and Planning A*, 10:1287–1305, 1978.

[Leonardi, 1980a] G. Leonardi. A unifying framework for public facility location problems – part 1: A critical overview and some unsolved problems. *Environment and Planning A*, 13:1001–1028, 1980.

[Leonardi, 1980b] G. Leonardi. A unifying framework for public facility location problems – part 2: Some new models and extensions. *Environment and Planning A*, 13:1085–1108, 1980.

[Leonardi, 1983] G. Leonardi. The use of random-utility theory in building location-allocation models. In J.-F. Thisse and H.G. Zoller, editors, *Locational Analysis of Public Facilities*, volume 31 of *Studies in Mathematical and Managerial Economics*. North-Holland, Amsterdam, 1983.

[Logan, 1985] B.I. Logan. Evaluating public policy costs in rural development planning: The example of health care in Sierra Leone. *Economic Geography*, 61:144–157, 1985.

[Love *et al.*, 1988] R.F. Love, J.G. Morris, and G.O. Wesolowsky. *Facilities Location: Methods and Models*. North-Holland, Amsterdam, 1988.

[Luenberger, 1984] D.G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA, second edition, 1984.

[Luoma and Palomäki, 1983] M. Luoma and M. Palomäki. A new theoretical gravity model and its application to a case with drastically changing mass. *Geographical Analysis*, 15:14–27, 1983.

[Malczewski and Ogryczak, 1995] J. Malczewski and W. Ogryczak. The multiplie criteria location problem: 1. a generalized network model and the set of efficient solutions. *Environment and Planning A*, 27:1931–1960, 1995.

[Martin and Bracken, 1991] D. Martin and I. Bracken. Techniques for modelling population-related raster databases. *Environment and Planning A*, 23:1069–1075, 1991.

[Martin and Williams, 1992] D. Martin and H.C.W.L. Williams. Market-area analysis and accessibility to primary health-care centres. *Environment and Planning A*, 24:1009–1019, 1992.

[Martin *et al.*, 1994] D. Martin, M.L. Senior, and H.C.W.L. Williams. On measures of deprivation and the spatial allocation of resources for primary health care. *Environment and Planning A*, 26:1911–1929, 1994.

[Martin, 1989] D. Martin. Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers, New Series*, 14:90–97, 1989.

[Massam and Malczewski, 1991] B.H. Massam and J. Malczewski. The location of health centres in a rural region using a decision support system: A Zambian case study. *Geography Research Forum*, 11:1–24, 1991.

[Mayhew and Leonardi, 1982] L.D. Mayhew and G. Leonardi. Equity, efficiency, and accessibility in urban and regional health-care systems. *Environment and Planning A*, 14:1479–1507, 1982.

[Mehretu, 1985] A. Mehretu. A spatial framework for redressing disparities in rural service delivery systems. *Tijdschrift voor economische en sociale geografie*, 76:363–373, 1985.

[Mehretua *et al.*, 1983] A. Mehretua, R.I. Wittick, and B.W. Pigozzi. Spatial design for basic needs in eastern Upper Volta. *Journal of Developing Areas*, 17:383–394, 1983.

[Mesa-Lago, 1985] C. Mesa-Lago. Health care in Costa Rica: Boom and crisis. *Social Science and Medicine*, 21:13–21, 1985.

[Minieka, 1970] E. Minieka. The *m*-center problem. *SIAM Review*, 12:138–139, 1970.

[Ministerio de Salud, 1991] Ministerio de Salud. Consultas de planificación familiar segun metodo. San José, Costa Rica, 1991.

[Moon and Chaudhry, 1984] I.D. Moon and S.S. Chaudhry. An analysis of network location problems with distance constraints. *Management Science*, 30:290–307, 1984.

[Moore and ReVelle, 1982] G.C. Moore and C. ReVelle. The hierarchial service location problem. *Management Science*, 28:775–780, 1982.

[Mosely, 1979] M.J. Mosely. *Accessibility: The Rural Challenge*. Methuen & Co, London, 1979.

[Mulligan, 1991] G.F. Mulligan. Equality measures and facility location. *Papers in Regional Science*, 70:345–365, 1991.

[Narula *et al.*, 1977] S.C. Narula, U.I. Ogbu, and H.S. Samuelsson. An algorithm for the *p*-median problem. *Operations Research*, 25:709–713, 1977.

[Narula, 1984] S.C. Narula. Hierarchical location-allocation problems: A classification scheme. *European Journal of Operational Research*, 15:93–99, 1984.

[Okabe *et al.*, 1992] A. Okabe, B. Boots, and K. Sugihara. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley and Sons, Chichester, 1992.

[Okafor, 1987] S.I. Okafor. Inequalities in the distribution of health care facilities in Nigeria. In R. Ahktar, editor, *Health and Disease in Tropical Africa*, pages 383–401. Harwood, London, 1987.

[Okafor, 1990] F.C. Okafor. The spatial dimensions of accessibility to general hospitals in rural Nigeria. *Socio-Economic Planning Science*, 24:295–306, 1990.

[O'Kelly, 1987] M.E. O'Kelly. Spatial interaction based location-allocation models. In A. Ghosh and G. Rushton, editors, *Spatial Analysis and Location-Allocation Models*, pages 302–326. Van Nostrand Reinhold, New York, 1987.

[Oliver and Webster, 1990] M.A. Oliver and R. Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, 4:313–332, 1990.

[Openshaw, 1973] S. Openshaw. Insoluble problems in shopping model calibration when the trip pattern is not known. *Regional Studies*, 7:367–371, 1973.

[Oppong and Hodgson, 1994] J.R. Oppong and M.J. Hodgson. Spatial accessibility of health care facilities in Suhum District, Ghana. *Professional Geographer*, 46:199–209, 1994.

[Oppong, 1992] J.R. Oppong. *Location-Allocation Models for Primary Health Care in Suhum District, Ghana*. PhD thesis, University of Alberta, Edmonton, Alberta, 1992.

[Paul, 1991] B.K. Paul. Family planning availability and contraceptive use in rural Bangladesh: An examination of the distance decay effect. *Socio-Economic Planning Science*, 25(4):268–282, 1991.

[Paul, 1992] B.K. Paul. Health search behavior of parents in rural Bangladesh: An empirical study. *Environment and Planning A*, 24:963–973, 1992.

[Phillips, 1986] D.R. Phillips. The demand and utilization of health services. In M. Pacione, editor, *Medical Geography: Progress and Prospect*. Croom Helm, London, 1986.

[Phillips, 1990] D.R. Phillips. *Health and Health Care in the Third World*. Longman Scientific, 1990.

[Poland *et al.*, 1990] B.D. Poland, S.M. Taylor, and M.V. Hayes. The ecology of health services utilization in Grenada, West Indies. *Social Science and Medicine*, 30:13–24, 1990.

[Rahman and Smith, 1991] S. Rahman and D.K. Smith. A comparison of two heuristics methods for the p-median problem with and without maximum distance constraints. *Internation Journal of Operations & Production Management*, 11(6):76–84, 1991.

[Reid *et al.*, 1986] R.A. Reid, K.L. Ruffing, and H.L. Smith. Managing medical supply logistics among health workers in Ecuador. *Social Science and Medicine*, 22:9–14, 1986.

[Rietveld, 1990] P. Rietveld. Infrastructure planning and rural development: Reflections on the urban functions approach. *International Regional Science Review*, 13:249–255, 1990.

[Ripley, 1981] B.D. Ripley. *Spatial Statistics*. John Wiley and Sons, New York, 1981.

[Rondinelli, 1985] D.A. Rondinelli. *Applied Methods of Regional Analysis*. Westview Press, Boulder, 1985.

[Rondinelli, 1990] D.A. Rondinelli. Location planning and regional development: Appropriate methods in developing countries. *International Regional Science Review*, 13:241–248, 1990.

[Rosero, 1993] L. Rosero. Physical accessibility to health facilities in Costa Rica. In *Proceedings of the International Population Conference*, pages 185–190, Liege, Belgium, August 1993. Ordina Editions.

[Rosero, 1995] L. Rosero. Spatial dimensions of family planning in Costa Rica: the value of geocoding demographic studies. Paper presented at the International Seminar on the Population of Central America, October 1995.

[Ross *et al.*, 1992] J.A. Ross, W.P. Mauldin, S.R. Green, and E.R. Cooke. *Family Planning and Child Survival Programs as Assessed in 1991*. The Population Council, New York, 1992.

[Rushton, 1984] G. Rushton. Use of location-allocation models for improving the geographical accessibility of rural services in developing countries. *International Regional Science Review*, 9:217–240, 1984.

[Rushton, 1988] G. Rushton. Location theory, location-allocation models and service development planning in the Third World. *Economic Geography*, 64:97–120, 1988.

[Rushton, 1993] G. Rushton. Lessons from the debate on location analysis in rural economic development. *International Regional Science Review*, 15:317–324, 1993.

[Segall, 1988] R.S. Segall. Mathematical modelling for the capacity planning of market oriented systems: with an application to real health data. *Applied Mathematical Modelling*, 12:366–378, 1988.

[Segall, 1989a] R.S. Segall. Some deterministic and stochastic nonlinear optimization modelling for the spatial allocation of multicategorical resources: with an application to real health data. *Applied Mathematical Modelling*, 13:641–650, 1989.

[Segall, 1989b] R.S. Segall. Some nonlinear optimization modelling for planning objectives of large market-oriented systems: with an application to real health data. *Applied Mathematical Modelling*, 13:203–214, 1989.

[Sibson, 1981] R. Sibson. A brief description of natural neighbour interpolation. In V. Barnett, editor, *Interpreting Multivariate Data*, pages 21–36. John Wiley and Sons, Chichester, 1981.

[Snickars and Weibull, 1977] F. Snickars and J. Weibull. A minimum information gain principle. *Regional Science and Urban Economics*, 7:137–168, 1977.

[Stevenson, 1987] D. Stevenson. Inequalities in the distribution of health care facilities in Sierra Leone. In R. Ahktar, editor, *Health and Disease in Tropical Africa*, pages 403–414. Harwood, London, 1987.

[Stimson, 1980] R.J. Stimson. Spatial aspects of epidemological phenomena and of the provision and utilization of health care services in Australia: a review of methodological problems and empirical analyses. *Environment and Planning A*, 12:881–907, 1980.

[Taket, 1989] A.R. Taket. Equity and access: Exploring the effects of hospital location on the population served – a case study in strategic planning. *Journal of the Operational Research Society*, 40:1001–1010, 1989.

[Tarimo, 1991] E. Tarimo. *Towards a Healthy District: Organizing and Managing District Health Systems based on Primary Health Care*. World Health Organization, Geneva, 1991.

[Teitz and Bart, 1968] M.B. Teitz and P. Bart. Heuristic methods for estimating the generalized vertex median of a weighted graph. *Operations Research*, 16:955–961, 1968.

[Tewari, 1992] V.K. Tewari. Improving access to services and facilities in developing countries. *International Regional Science Review*, 15:25–37, 1992.

[Tien and El-Tell, 1984] J.M. Tien and K. El-Tell. A quasihierarchial location-allocation model for primary health care planning. *IEEE Transactions on Systems, Man, and Cybernetics*, 14:373–380, 1984.

[Tobler, 1983] W. Tobler. An alternative formulation for spatial-interaction modeling. *Environment and Planning A*, 15:693–703, 1983.

[Toregas *et al.*, 1971] C. Toregas, R. Swain, C. ReVelle, and L. Bergman. The location of emergency services. *Operations Research*, 19:1363–1373, 1971.

[Tsui and Ochoa, 1992] A.O. Tsui and L.H. Ochoa. Service proximity as a determinant of contraceptive behaviour: Evidence from cross-national studies of survey data. In J.F. Phillips and J.A. Ross, editors, *Family Planning Programmes and Fertility*, pages 222–256. Clarendon Press, Oxford, 1992.

[Tsui *et al.*, 1981] A.O. Tsui, D.P. Hogan, J.D. Teachman, and C. Welti-Chanes. Community availibility of contraceptives and family limitation. *Demography*, 18:615–625, 1981.

[Tsui, 1982] A.O. Tsui. Contraceptive availibility and family limitation in Mexico and rural Korea. *International Family Planning Perspectives*, 8:8–21, 1982.

[Webber and O'Kelly, 1981] M.J. Webber and M.E. O'Kelly. Emprical tests and sensitivity analysis of a model of residential and facility location. *Geographical Analysis*, 13:398–411, 1981.

[Webber, 1979] M.J. Webber. *Information Theory and Urban Spatial Structure*. Croom Helm, London, 1979.

[Weibull, 1976] J.W. Weibull. An axiomatic approach to the measurement of accessibility. *Regional Science and Urban Economics*, 6:357–379, 1976.

[Weibull, 1980] J.W. Weibull. On the numerical measurement of accessibility. *Environment and Planning A*, 12:53–67, 1980.

[WHO and UNICEF, 1978] *Primary Health Care: Alma-Ata 1978*. World Health Organization and United Nations Children's Fund, Geneva, 1978.

[WHO, 1981] *Global Strategy for Health for All by the Year 2000*. World Health Organization, Geneva, 1981.

[WHO, 1992] *Reproductive Health: A Key to a Brighter Future*. World Health Organization, Geneva, 1992.

[WHO, 1994] *Information Support for New Public Health Action at District Level*. Number 845 in WHO Technical Report Series. World Health Organization, Geneva, 1994.

[Wilson and Gibberd, 1990] R.M. Wilson and R.W. Gibberd. Combining multiple criteria for regional resource allocations in health care systems. *Mathematical Computer Modelling*, 13(8):15–27, 1990.

[Wilson *et al.*, 1981] A.G. Wilson, J.D. Coelho, S.M. Macgill, and H.C.W.L. Williams. *Optimization in Locational and Tranport Analysis*. Wiley, Chichester, 1981.

[Wilson, 1974] A.G. Wilson. *Urban & Regional Models in Geography & Planning*. Wiley, New York, 1974.

[World Bank, 1993] *Effective Family Planning Services*. The World Bank, Washington, D.C., 1993.