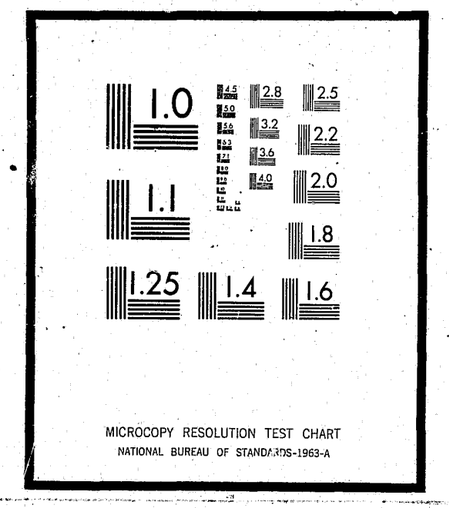


OPERATIONAL ASPECTS OF EMERGENCY AMBULANCE SERVICES

by

KEITH ALLISTER STEVENSON

Technical Report No. 61

OPERATIONS RESEARCH CENTER

MASSACHUSETTS INSTITUTE
OF
TECHNOLOGY

May 1971

1/13/25

OPERATIONAL ASPECTS OF EMERGENCY AMBULANCE SERVICES

by

KEITH ALLISTER STEVENSON

Technical Report No. 61

Work Performed Under

National Science Foundation Grants
Operations Research for Public Systems
Grant GK-1685 MIT/DSR 70546
Grant GK-16471 MIT/DSR 72326

Operations Research Center
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

May 1971

Adapted from a thesis, supervised by Professor Alvin W. Drake, presented to the Department of Electrical Engineering in partial fulfillment of the requirements for the degrees of Master of Science and Electrical Engineer, June, 1970.

FOREWORD

The Operations Research Center at the Massachusetts Institute of Technology is an interdepartmental activity devoted to graduate education and research in the field of operations research. The work of the Center is supported, in part, by government contracts and industrial grants-in-aid. The work reported herein, its supervision, and other expenditures associated with it were supported by the National Science Foundation, under Grants GK-1685 and GK-16471.

John D. C. Little
Director

ACKNOWLEDGEMENT

I should like to express my gratitude to Professor Alvin W. Drake of the M.I.T. Operations Research Center who supervised this research. His guidance, encouragement and patience are all sincerely appreciated.

A special note of thanks is due to Dick Larson who shared his interest, advice and friendship unstintingly, and gave freely of his time to read and reread drafts of this document. Thanks are also due to John Jennings and Ralph Keeney for time spent draft-reading and formulating helpful suggestions.

I want also to thank Doug Lazarus of the N.Y.C.-Rand Institute for his art work, and Mrs. Constance Hood for her superb typing.

Finally, I acknowledge gratefully the financial support of the National Science Foundation, and the moral support of the M.I.T. Operations Research Center.

Keith A. Stevenson

ABSTRACT

Despite the recent appearance of a number of reports and their recommendations on improving emergency ambulance systems, there is reason to believe that the actual changes and improvements in most large urban operations are small. A major problem is the difficulty encountered by an administrator in trying to translate general recommendations into specific proposals with regard to the type and the size of the operation with which he is concerned. This document attempts to provide a rational framework for decision-making so that an administrator may initiate changes with some degree of confidence.

The analysis of the emergency ambulance service is based on the importance of the time elapsing between the occurrence of the emergency and the first arrival of a properly equipped ambulance and its well trained crew. This response time is considered in terms of two components: the dispatch delay (the time between the reception of the call and an ambulance being available to respond) and the travel delay (the time taken for the ambulance to travel to the scene of the emergency). As a first approach to modelling the dispatch delay simple results from the theory of queues are invoked to relate the probability of a dispatch delay and its expected length to the rate at which calls arrive, the time taken to service calls, and the number of vehicles assigned to the service. As a result of the insights afforded by this modelling effort, another model is developed to show how a higher quality, reduced cost service may be provided using "secondary" ambulances. By making simplifying assumptions about the spatial distribution of demand and the street layout in a city, models of the expected travel delay are developed for different distributions of ambulances.

Using estimates of the costs of vehicles, equipment and manpower it becomes possible for the ambulance service administrator to compare the level of service, in terms of the response time, with the expected costs of providing such a service. This capability is important in the design of emergency ambulance systems, in predicting budget requirements in the future, and evaluating alternative methods for providing the service.

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

TABLE OF CONTENTS (Cont.)

LIST OF FIGURES

LIST OF FIGURES (Cont.)

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

1.1 Background

For many centuries men have pursued wars with great earnestness and concern. It was not, however, until the Napoleonic Wars of the early nineteenth century that history was able to record an interest in the well-being of the soldiers who were wounded in these bloody conflicts. Baron Jean Larrey, Napoleon's Surgeon-General is the man to whom credit must be given for first organizing the transport of the injured to facilities located away from the battle, where medical treatment was administered.[10]*

Ambulance transportation first came to America through another soldier, Dr. Edward B. Dalton, who arranged for the transport of injured members of the Army of the Potomac during the Civil War. Not surprisingly, perhaps, emergency ambulance transportation has retained a low-priority position in our present highly developed, urbanized society. Men have always been more concerned with the relatively glamorous public service tasks of apprehending criminals and extinguishing fires, than with transferring sick or injured people to hospitals. The result is that while cities and towns across the country have tightly organized police forces and fire departments with their relatively well paid members, the provision of emergency medical transportation has been left

---

*Numbers identify references listed at the end of the document.

to the consciences of civic-minded people. Consequently the provision of emergency ambulance service across the country appears to be a disorganized and haphazard potpouri of people and interests. The list of organizations providing such a service includes morticians and small private operations that often struggle to stay in business, municipal police and fire departments that include ambulance services as an adjunct to their other work, small groups of volunteers, and such unlikely candidates as taxi-cab fleets and gasoline stations. Among many of these groups the quality of the care and transportation leaves much to be desired, not necessarily because of obtuseness, but often because of a lack of information and finance.

In the chaotic picture of ambulance services, there are discernable trends to which we shall return in a later chapter. These are the results of changes in the political and legal environment in which the services operate. These changes have developed following a great surge of interest in the subject over the last four or five years. For more than two decades the medical fraternity has complained in its journals about the state of emergency health services in the nation, and about the emergency ambulance services in particular. It was not until the publication in 1966 of "Accidental Deaths: The Neglected Disease of Modern Society"[1] by the National Academy of Sciences and the National Research Council that the concern over poor standards of both equipment and attendants' training became nationwide. This monograph pointed out that deaths from accidents were rising at an alarming rate, especially those in

automobile accidents. We reproduce some relevant statistics in the table below:[16]

Table 1.1: Deaths from Accidents in the U.S.A.

	Deaths (1000)			Deaths per 100,000 population		
	1950	1960	1966	1950	1960	1966
All accidents	91.2	93.8	113.6	60.6	52.3	58.0
Motor vehicle accidents	34.8	38.1	53.0	23.1	21.3	27.1

The link between an accident victim and the hospital in which his hopes for survival may lie is the emergency ambulance and its crew. The importance of what had been a neglected service received attention for the first time.

There were other changes in our society that have helped to focus attention on the state of emergency ambulance services. The rising costs of medical care are too well known to require documentation. In addition there has been a steady decline in the number of physicians in general practice, as is indicated in the following table:[16]

Table 1.2: Physicians in the USA: Number Per 100,000 Population

	1950	1955	1960	1965	1967
Active Physicians	141	141	140	145	150
Private Physicians	109	102	98	100	100
General Practitioners	72	59	47	39	36

The net result is that many people can neither find nor afford a private doctor. Instead of receiving treatment at home, they must travel to the emergency ward or the outpatients' department at a local hospital. In cases of sudden severe illness or accident, these people may need to use an emergency ambulance service. The steady increase in the use of these services as documented extensively in the report, "Description and Analysis of 18 Proven Highway Emergency Medical Systems" by Cooper et al.[2] confirms the above hypothesis, as do the overworked ambulatory service facilities in hospitals all over the country.[3,14,15,22,23]

Finally, we note the warning sounded by Dr. David Rutstein in his book, "The Coming Revolution in Medicine,"[18] that the trend towards the regionalization of medical care can only be accomplished with "careful planning and an efficient transportation and communications system." Under regionalized hospital systems, there will be an increase in the volume of inter-hospital traffic. In addition, with hospitals specializing in particular types of treatment, ambulances will have to be dispatched to emergency wards that are appropriate to the condition of the individual patient.

1.2 Emergency Ambulance Operations

As a first step towards understanding more clearly why problems exist in the provision of ambulance service, we shall briefly review the different methods of operation. Since there exists no national coordinating body to supervise the organization of services, the variety encountered is not really surprising. We may divide the different operations into four broad groups:

(1) Private concerns

(2) Volunteer groups

(3) Municipal services

(4) Combinations of the above.

1.2.1 Private Concerns

According to a survey reported in reference [7], until about five years ago 91 percent of the municipal operations in the U.S. were in the hands of private companies. Of these about 80 percent were funeral homes. The survey was taken over cities of very different sizes, the majority of which had populations of less than 25,000, and it is therefore difficult to draw strong conclusions; but it is known that until very recently morticians operated nearly one-half of the nation's 40,000 ambulances.

The same survey indicates that funeral homes are the major emergency ambulance purveyor in the southern states of the country, while private companies (non-mortician) are by far the most prominent on the West Coast. Morticians and private companies generally provide the emergency service in small towns, usually as an adjunct to the regular, more routine ambulance transfers that form the major part of the private operation's business. Two very large private emergency services are in Seattle, Washington and Dade County, Florida.

1.2.2 Volunteer Groups

In small towns, rural areas and even some of the burgeoning suburbs, members of the community have banded together to provide a

service staffed entirely by volunteers. By removing salaries from the expenses of operation, it has been possible to provide emergency transport where economically it appeared infeasible. Funds are usually collected from the community by means of charity drives.

As wealthy an area as Bethesda-Chevy Chase in Maryland, for example, is served by a volunteer service.

1.2.3 Municipal Services

In most of the large cities, New York, Washington, Chicago, San Francisco, Boston, Baltimore, the emergency ambulance service is provided by one or another municipal agency, usually free of charge to the patient. The types of agency can conveniently be broken into three groups:

(a) Police: Station wagons and paddy wagons used for patrol duties have been equipped with bandages, splints, oxygen tanks, and stretchers, and respond to requests for emergency transportation received by the police dispatcher. Since the police vehicles are on continuous patrol throughout the city, the response time (i.e., the time between the reception of a call and the arrival of an ambulance) to an emergency is very small. The cost of this type of system is low, because time not spent on ambulance duty is spent on preventive patrol. The disadvantages of the operation are that the vehicles do not meet current state specifications as ambulances, and that the first aid training received by policemen is minimal. Two

cities providing the service in this manner are Boston and Louisville, Kentucky.

(b) Fire Departments: Firemen have always received first aid as a part of their training, and most fire departments have rescue companies (or squads) whose task it is to rescue and treat inhabitants and firemen injured in burning buildings. The rescue squads use vehicles designed to transport patients and equipped with first aid supplies. In small cities the fire department therefore often provides the non-fire emergency service. A typical example is the city of Cambridge, Massachusetts. In other instances the fire department has been the nucleus around which a volunteer ambulance service has formed. In Chicago and Baltimore the fire departments have formed a separate squad to provide city-wide emergency service.

(c) Hospitals: Despite their close associations with the ambulance service there appear to be relatively few hospitals that provide an emergency service. Some hospitals do have their own ambulances, but one of the few cities with a hospital based emergency ambulance service is New York. Ambulances located at hospitals throughout the city respond to calls in regions associated with each hospital. Patients are usually taken to the hospital from which the ambulance was dispatched. An advantage of this system is the possibility of the attendants receiving training at the hospital, and perhaps of working in

the emergency ward. A disadvantage is the long response time from hospitals serving large regions because the ambulances are not dispersed geographically.

1.2.4 Combinations

It is not difficult to imagine the possible combinations of the systems described above. In almost any city, the police by the nature of their work, play some part in the emergency service. Very often the police department dispatches the ambulances, and police vehicles, even when not the primary emergency service, transport people with minor injuries to hospitals. In the city of Cambridge, Mass., the police department receives calls for transportation, and also handles minor injuries. The fire department provides the major emergency service, with support from a private company. In one Rhode Island community, Pawtucket, the fire rescue squad answers calls, administers treatment and then turns the patient over to a private company for transport to the hospital.

1.3 Recent Developments

The first response to the recent appeals for improved ambulance services came with the passage of state laws specifying minimum standards of operation for those engaged in the service. These covered the training and licensing of ambulance personnel, the equipment carried on the ambulance and some features of the vehicle's design. A typical example of these standards is included in this document as Appendix C.

A significant step occurred in 1966 when ambulance personnel were included in the provisions of the Fair Labor Standards Act, and thereby qualified for a minimum wage of $1.60 an hour. Previously, operators running an ambulance business that was only marginally profitable kept salaries below the minimum wage. An obvious result was that the people employed as ambulance attendants were those who were unable to find other, more lucrative work, and therefore not necessarily the most suitable for the job.

The Department of Transportation, which is the Federal agency most closely connected with improving the quality of emergency transportation, has funded a number of studies. These have included issuing guidelines for the operation of services, analysis of existing operations, suggested training programs, and fairly detailed economic analyses.

Finally, the extension of the Medicare and Medicaid programs to include medically necessary transportation has eased the financial burden on some ambulance operators, and allowed the improvement of equipment and the increase in salaries. In addition, under the Highway Safety Act (1966), some money has been made available to subsidize the provision of emergency services related to highway accidents.[17]

1.4 Current Concerns

The passage of high-minded legislation, while it may connote concern about a social problem, does not necessarily reflect complete understanding of the causes of the problem nor an intention to provide the means to overcome these. The legislation defining minimum standards

of ambulance practice, while a necessary step in improving the quality of emergency care, did not provide the means for carrying out the improvements. Many small private companies operating their emergency services close to the margin have simply declined to continue this aspect of their work when faced with the prospect of increased costs to meet higher standards. Even in some of our larger cities the emergency ambulance service is provided as it always was at standards that are now legally unacceptable because no money is forthcoming to carry out improvements.

Even where money has been available (under the Highway Safety Act, for example) it has been left largely untouched because of the difficulty encountered by administrators in trying to translate general recommendations into specific proposals with regard to the type and size of the operation with which they are concerned. One of the main concerns of this report is an attempt to provide a rational framework for decision-making so that the administrator might initiate changes with some degree of confidence.

In the nationwide survey reported in reference no. 5, the following were the major problems outlined by administrators of current ambulance systems (in order of importance):

(a) The hiring and retention of reliable personnel: Low salaries and poor working conditions combine to discourage most people from a job that involves great personal responsibility and emotional stress. A consequence has been that very often people employed as ambulance attendants are those least qualified to react appropriately in an emergency situation.

(b) Finances: Collecting payment for ambulance transport in an emergency situation appears to be a problem in most commercial operations. At the time of the emergency it may be impossible or improper to request payment, so that the operator must rely on postal billing. Because the patient's bill reflects the average rather than the marginal cost of his trip* it usually appears to be excessively large in relation to the services provided. Consequently private operations expend a great deal of effort trying, often unsuccessfully, to collect payment.

(c) Travel distances and response times: These are often great, either because there are too few vehicles allocated to an area or because the vehicles are poorly located. Some of the models presented in this document should help administrators to overcome this problem.

(d) Communications and coordination: The emergency ambulance communication and coordination problems present in most communities exist because there is no individual or group of individuals responsible for the operation of the system. Radio dispatching ambulances and notifying emergency wards of impending arrivals are within the capabilities of most systems, yet they are functions often poorly performed. The coordination of the different emergency systems within a community is a more difficult problem, as is the task of inter-community

*I.e., the patient pays the cost of having ambulances available to respond immediately to emergencies in his community.

coordination. These are unlikely to be solved until there is some central responsibility for their solution.

From the patient's point of view the problems are very similar. The perceived delay until emergency assistance arrives is usually enlarged for the victim of an emergency and his relatives. Where this delay is significant from an objective point of view, it will appear to be enormous to the patient. The quality of the care administered by untrained individuals is unlikely to be ignored by the victim in an emergency who is very aware of the consequences of mismanagement. Finally, as we mentioned before, the financial burden that the patient bears may seem inequitable, for the patient pays not the cost of the ambulance and its crew undertaking the journey, but a cost associated with always having emergency service available to every member of the community. Most of our large cities have adopted the liberal view that, as with fire and police service, the community should spread the risk of emergencies by sharing the costs associated with providing ambulance service.

1.5 Related Research

Until 1966 most of the research into the problem of emergency ambulance services was confined to reports enumerating the shortcomings of operations in particular cities, and to the private research of concerned doctors published in medical journals. In that year the National Academy of Sciences published "Accidental Death and Disability: The Neglected Disease of Modern Society."[1] In 1967 Manegold and Silver published an article[14] in the Journal of the American Medical Association

in which they qualitatively discussed the emergency medical "system." The article specifically isolated the delay between the incident and the arrival of an ambulance as an important factor, and tried to formulate the quality of service in system terms. At the end of 1967 David C. Dimendberg of the New York City Department of Hospitals published the results of a month-long survey of ambulance transport in the city.[3] Using some of the results of that survey Emmanuel Savas undertook a detailed study of the King's County Hospital district in Brooklyn. Using a computer simulation model Savas demonstrated that the average response time and the average service time could both be reduced considerably if some of the ambulances stationed at the Kings County Hospital were transferred to a garage located in an area in the district in which the call density was extremely high. Additional recommendations from this report were that the emergency service be separated from the hospitals, that ambulances be as widely dispersed as possible and that dispatching be from one central point directly to the ambulance without any intermediaries.[21]

In 1968 Dunlap and Associates issued the results of a study of the economics of emergency ambulance transportation.[5] The report included the results of the first nationwide survey of emergency ambulance services, and it indicated changes in the type of purveyor (specifically a decrease in the number of morticians) and the problems experienced by operators. The Dunlap report used a very simple queuing model to predict the availability of ambulances, and suggested some simple procedures for locating ambulances. Both of these will be considerably expanded in

this document. The Dunlap study indicated the availability of funding to aid communities in establishing emergency ambulance facilities; and attempted an analysis of the efficiency of helicopter ambulances.

1.6 Scope of this Document

In this document we develop simple analytical models that describe some parts of the emergency ambulance system, and that can be used to improve its design and operation. In the previous section we outlined the simulation modelling work of Savas in New York City. We have chosen to tackle a similar problem by analytical means because computer simulations are expensive undertakings, especially when they involve a great deal of data-gathering. Furthermore, detailed simulations of one locality may require extensive data-collection and program design before being applied to another location. Simple analytical models, while providing less precision, can give approximate results that are applicable in a wide variety of communities. In most instances these models offer insights that can reduce the computations in cases where simulation is the only alternative.

We begin by isolating the transportation subsystem of the total emergency ambulance service system and identifying the relevant elements that bear on total system performance. We note that as in many public systems the output of the total system, the patient's condition, is not easily measured, and that surrogate measures need to be found. One of these, the time that a patient waits between the occurrence of the emergency and the arrival of the ambulance, is believed to be particularly crucial. The waiting time is reduced to its three components:

(i) the time until the emergency ambulance service is notified of the emergency

(ii) the time until an ambulance becomes available to respond to the emergency

(iii) the time that the ambulance takes to travel to the scene.

Since (ii) and (iii) are within the compass of the emergency ambulance service, it is on these that we concentrate, identifying their sum as the response time.

In the first part of Chapter 4 we develop a queuing model that allows the determination of the probability that an ambulance is unavailable, the expected delay before the dispatch of an ambulance, and the ambulance utilization in terms of the average rate at which emergency calls arrive, the time taken for an ambulance to complete such a call, and the number of ambulances in the service. In the second part of Chapter 4 we calculate the approximate expected time for an ambulance to travel to the scene of the emergency. By quoting an important result on the expected distance between an incident and randomly distributed vehicles due to Larson[13], we can estimate an upper bound on the time for the nearest available ambulance to get to an emergency. By using the dispatch delay time and the travel time models in combination, we can demonstrate a procedure to allocate a given number of ambulances to a region so that the expected response time is minimized.

One of the insights into emergency ambulance operation afforded by the models in Chapter 4 is the low utilization of resources in small operations. Chapter 5 is devoted to exploring possible methods of reducing costs, and includes a model of the situation in which a secondary source of ambulances is introduced in support of the primary source. From

the analysis it appears that a service with a highly paid secondary source can operate at a lower cost and at a higher level in terms of ambulance availability than a single source.

Two problems that are of related interest have been relegated to the appendices. In Appendix A we develop the probability law describing the total "service time" of an ambulance responding to emergency calls. The service time is defined as the sum of the times for the ambulance to travel to the scene, to administer treatment at the scene, and to transport the patient to hospital. Because of its importance in our models, we need to refine our estimates of the service time probability distributions if we are to improve the accuracy with which the models reflect reality. The effort seems worthwhile when one recalls that other municipal service operations encounter similar service time configurations.

In Appendix B we develop a simple graphical algorithm for directing ambulances to the hospital nearest the scene of the emergency. The procedure involves describing a region around each hospital so that an incident located inside hospital X's region is closer to that hospital than any other. There is a surprising need for systematizing such a procedure, since patients are often taken to distant hospitals according to the whims and habits of the ambulance personnel.

Finally, Appendix C contains a set of specifications which the State of Massachusetts requires to be observed by ambulance operators in the state. It is included as an example of the legislation slowly being enacted across the country. While its effect is probably to prompt improvements in operating standards, until the means for financing improvements are made available, ambulance services will operate in

violation of these specifications or else will close down.

In summary, this document approaches the problems of providing emergency ambulance service analytically, through the development of mathematical models. These models are the "outputs" of the study, and they can be used in the following ways:

(i) As guidelines to ambulance service administrators in allocating their resources (vehicles and men) among different shifts in various parts of the city or town being served.

(ii) As means for predicting resource requirements for administrators designing new services or improving old ones, and attempting to appropriate funding from the municipal budget.

(iii) As sources of insight into the way an emergency system operates, how the various elements interact, and therefore as guides to policy-making.

(iv) As aids to future modelling or simulation efforts undertaken to better achieve the above.

(v) In modelling other municipal service systems like the police, fire and sanitation services.

The reader will be reminded constantly in the succeeding chapters that analytical models, while providing a framework within which to approach the problem, are based on simplifying assumptions about system functioning. In practice, an ambulance operation faces a host of political, bureaucratic and labor constraints that may override, or at least compromise, the solutions suggested by modelling efforts.

CHAPTER 2: SOME STATISTICS OF EMERGENCY AMBULANCE OPERATIONS

2.1 Introduction

Chapter 1 gave some historical perspective to the problems of successfully providing emergency ambulance service, and gave a general indication of current conditions. Gaining specific information and "hard" data about emergency ambulance service is particularly difficult. This is a problem shared by many municipal services, and it is compounded in our case by the fact that accounting procedures differ among different systems according to who provides the service. For example, if the police provide the emergency ambulance service, every call for police service that has an emergency ward as its final disposition might be counted as an emergency call. On the other hand, when some other agency provides the emergency service, emergency ambulance-type transport carried out by the police is often ignored. Then again, when the emergency ambulance service is hospital-based, the tally of emergency calls may include a significant number of routine patient transfers (hospital-to-home and vice-versa, or inter-hospital), which are usually undertaken by the private ambulance companies.

The purpose of this short chapter will be to cull from the little data that is available descriptors of the emergency ambulance service*

*The reader is reminded that throughout this document we are concerned about emergency ambulance service. We shall not consider the routine transfer of patients to and from hospital in-patient facilities in this category.

that will provide details of real operations and facilitate their comparison. What follows has been accumulated out of observations and studies by the author in the Greater Boston area, and reports in the literature, especially two issued by the U.S. Department of Transportation: "Economics of Highway Emergency Ambulance Services" (Reference No. 5) and "Description and Analysis of 18 Proven Highway Emergency Medical Systems" (Reference No. 2).

2.2 The Provision of Emergency Ambulance Service

In 1967 the Dunlap study group[5] undertook the first extensive nationwide survey of emergency ambulance services. The survey received responses from 1763 municipally-based and 535 county-based services. We shall draw on the results of the survey to indicate the distribution of emergency ambulance service by the type of operator, and to support some of the remarks made in Chapter 1 about current trends in service provision. The survey is very heavily biased toward operations serving populations of less than 25,000, which explains the prominence of the volunteer squads in Table 2.1 below. In that table we indicate the distribution of service by purveyor. We note the continued importance of the mortician in the emergency ambulance service, and the very small role played nationwide by hospital-based emergency ambulances.* It is apparent from other information received in the survey that the mortician purveyors are mainly concentrated in the Southern States, that private

*New York City, which provides the largest service in the country uses hospital-based emergency ambulances.

Table 2.1 Distribution of Emergency Ambulance Operations by Type

Type of Purveyor	Percentage of the Municipal Services Surveyed	Percentage of the County Services Surveyed
Funeral Home	27.8	34.7
Private Company	23.8	15.1
Volunteer Squad	24.0	26.0
Local Government Agency	20.2	14.8
Hospital	3.7	8.0
Other	0.5	1.4

companies are most prominent on the West Coast, that local government agencies predominate in the North-Eastern part of the country, and that volunteer squads are generally widespread and confined to small services.

In approximately 10% of both the municipal and country respondents to the survey, there had been a change in the type of purveyor within the previous 5 years. For those services in which a change had occurred, we show in Table 2.2 the distributions of the type of purveyor, both prior to and following the change. From the above, there are a number of discernable trends in the way in which emergency service is changing:

(i) A definite and drastic movement away from the mortician as a source of emergency ambulance service.

(ii) An increase in the number of local government agencies taking responsibility for providing emergency ambulance service.

(iii) An increase in the number of volunteer squads in operation, possibly the result of disillusionment with morticians, but

Table 2.2 Distribution of Emergency Ambulance Operations by Type Prior to and Following a Change in the Purveyor

Type of Purveyor	Percentage of the Municipal Services Surveyed		Percentage of the County Services Surveyed	
	Formerly	Currently	Formerly	Currently
Funeral Home	72.5	2.6	78.3	1.7
Private Company	19.1	43.3	16.7	18.3
Local Government Agency	5.6	21.3	5.0	25.0
Volunteer Squad	0.6	13.4	0.0	13.3
Hospital	1.1	4.4	0.0	6.7
Combination	0.0	12.4	0.0	35.0
Other	1.1	2.6	0.0	0.0

probably because a volunteer squad was perceived as the only alternative to a small community when a mortician decided to discontinue his emergency ambulance service for financial reasons.

(iv) An increase in the number of municipal services employing private companies. This is probably a reflection of changes on the West Coast where the private company is a favored type of purveyor.

(v) No real indication that hospitals are becoming involved in the emergency ambulance service.

2.3 The Generation of Emergency Ambulance Patients

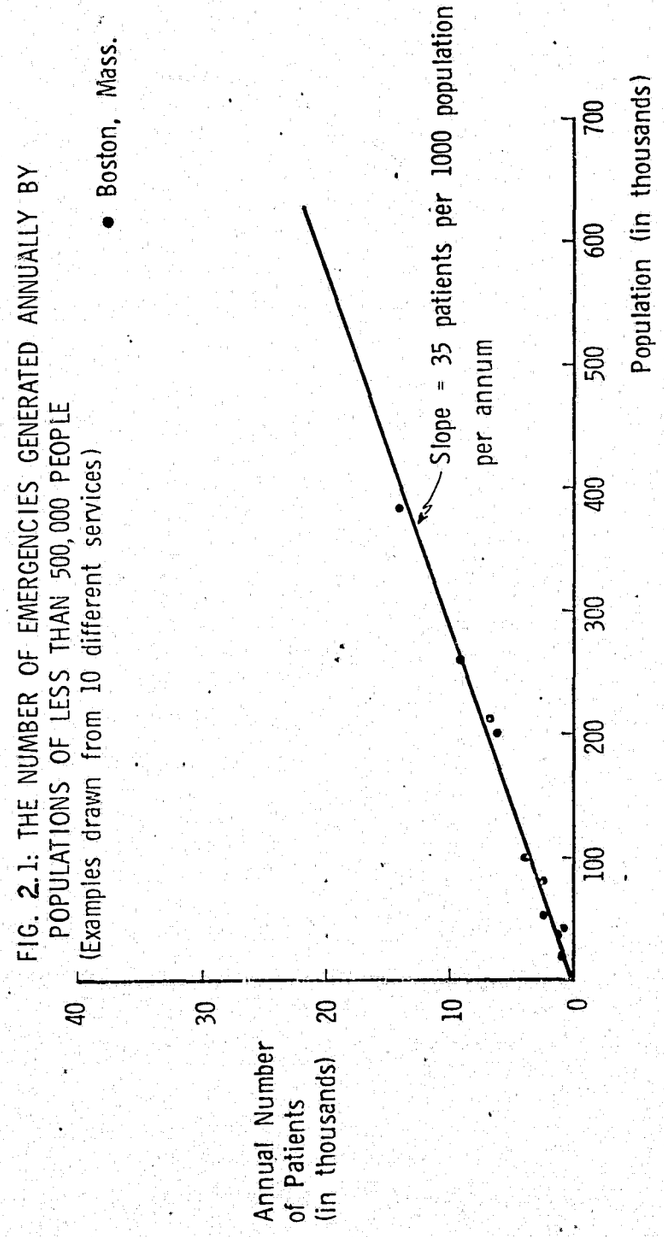
A commonly quoted rule of thumb among ambulance operators is the statistic: 35 emergency ambulance patients generated per thousand population per annum. While this is a useful indicator of the number of emergency ambulance calls that we might expect annually from a city or region, it is very approximate. From the little information available to the author, the statistic seems to hold up reasonably well for populations below 500,000 (see Figure 2.1). For larger populations the statistic is much less consistent and reflects socio-economic factors and the difference in accounting procedures mentioned in Section 2.1. Ideally (i.e., under conditions where patient accounting procedures were identical everywhere) the expected number of ambulance patients per thousand population per annum could be used:

(i) To predict emergency ambulance needs in the face of changing population.

(ii) To compare different cities or regions and to investigate the causes of extreme deviations in the number of patients generated per 1000 people.

Unfortunately these ideal conditions do not exist, and while the comparison of the above statistic among different emergency ambulance services may be interesting and informative, it does not justify the popular conclusion that services with a large number of emergency ambulance patients per 1000 population are being misused.

Compounding the above is the steady annual growth in the number of emergency ambulance calls generated. This rate exceeds the population



growth rate and requires constant expansion of the emergency ambulance service. In Table 2.3 we indicate the annual number of calls answered by three very different emergency ambulance services.[2,22]

Table 2.3 Annual Number of Calls Answered by 3 Emergency Ambulance Services

Service	1962	1963	1964	1965	1966	1967
Boston, Mass. (Police)	27,834	29,274	29,211	31,658	34,035	37,068
Baltimore, Md. (Fire)	-	37,597	37,594	39,185	40,000	41,910
Chevy Chase, Md. (Volunteer)	5,478	5,719	5,832	5,900	6,008	-

The causes of these increases are numerous. There has been a rise in the number of highway accidents, which often constitute 25% or more of the emergency ambulance load. The decline of the general practitioner has driven a lot of patients to municipal emergency wards, especially among the poor. Urban civil disturbances over the last five years have made notable additions to the emergency ambulance load.

2.4 The Costs of Providing an Emergency Ambulance Service

The cost of providing an emergency ambulance service is a function of the type of service, the location and the quality. In the next section we shall give a few specific examples. In this part of the chapter we shall give a general outline of the distribution of costs among the various functions in the service. In Table 2.4 we indicate

as a percentage of the total budget, the proportions for which different aspects of the operation are responsible.*

Table 2.4 The Distribution of Emergency Ambulance Costs

Item	Percentage of Total Budget
Attendant Salaries and Benefits	65 - 80
Support Personnel Salaries	10 - 15
Insurance	1 - 2
Equipment Depreciation	5 - 7
Repairs and Maintenance	2 - 3
Laundry	1
Station Rental and Operation	3 - 5

The notable aspect of this breakdown is the very high fraction of the total cost that is due to attendant salaries. In general, municipal operators reported attendent salaries at the upper end of the range, while private operators in rural areas reported salaries at the lower end.[5] With the outlook being toward higher rather than lower salaries for attendents, attendents will in the future be responsible for an even larger fraction of total cost. Under these circumstances we shall feel justified in Chapter 5 in approximating emergency ambulance costs as a linear function of the number of ambulances manned during the day.

*These proportions reflect the results of the surveys carried out in References 2 and 5.

2.5 Some General Descriptive Statistics

From the general data published by emergency ambulance services about their operations, it is possible to derive a few informative statistics. Because of the way the data is collected, these statistics do not make reasonable measures of system performance. Instead they are essentially descriptive, and at best give insights into general operating procedures. The following is a list of the more useful of the statistics that can be easily obtained from the aggregated data of most emergency ambulance services:

(i) The number of emergency calls per thousand population per annum.

(ii) The cost per patient transported.

(iii) The cost per resident per annum .

(iv) The cost per ambulance per annum .

(v) The number of patients per ambulance per day.

All of the above give a general indication of current loads and costs, and will provide guidelines for some of the more theoretical work of future chapters. (v) above is indicative of the very low utilization of ambulances in current operations. In Table 2.5 we provide examples of these statistics for 6 different cities, with very different characteristics and different types of emergency ambulance service.[2,3,22]

The reader is warned against drawing dramatic conclusions from Table 2.5. Instead his attention is drawn to the large variation in the statistics we have identified, the generally high cost per patient, the relatively low annual cost per resident, and the low utilization of the

Table 2.5 General Descriptive Statistics for 6 Cities

City	Population	No. of Ambulances	No. of Ambulance Runs per Annum	No. of Ambulance Runs per 1000 Pop.	Cost of Service Per Annum	Cost per Patient	Cost per Resident per Annum	Cost per Ambulance per Annum	Patients per Ambulance per Day
	(thousands)		(thousands)		(thousands)	($)	($)	($ thousands)	
New York (Hospital)	8,000	110	550	69	7,000	13	0.9	63.6	13.7
San Francisco (Municipal)	952	16	17	18	1,100	65	1.2	69.0	2.9
Baltimore (Fire)	939	14	42	43	915	21	1.0	65.0	8.2
Boston (Police)	618	40	37	60	250	7	0.4	6.3	2.6
Cambridge Mass. (Fire & Police)	100	7	3.8	38	132	35	1.3	18.8	1.5
Petersburg Virginia (Private)	38	2	1.2	32	43.8	37	1.2	20.0	1.6

ambulances (patients per ambulance per day). Not surprisingly the cost per patient in Boston, the only city in which the service is the exclusive province of the police department, is very low. This does not necessarily imply that all emergency ambulance services should be given to the police, since they would be the first to admit that it is extremely difficult for policemen and police vehicles to meet the rigorous standards required for a high quality ambulance service.

2.6 Chapter Conclusion

In this chapter we have quantitatively outlined current trends in the provision of emergency ambulance service. We have discussed the generation of patients, and explored, with examples, some of the more readily available statistics relating to emergency ambulance operations. With these as guides we proceed in Chapter 3 to structure emergency ambulance service as a system, and then in Chapters 4 and 5 to model parts of the system.

CHAPTER 3: THE EMERGENCY AMBULANCE SERVICE SYSTEM

3.1 Introduction

In this chapter we consider the emergency ambulance service as a system, and thereby provide the framework within which some aspects of the service may be rationally analyzed. Very simply, a system consists of a number of components or elements, all hopefully directed toward the same purpose, which operate on a set of inputs to produce a set of outputs. Characteristically the interaction of the elements is complex: altering one element in the system affects the other elements in a way that is often not straightforward. Typically the process of systems analysis entails the definition of the objectives of the system, the generation of alternative ways of meeting the objectives, the definition of measures with which to evaluate the alternatives, and the ultimate selection of the "best" alternative.

We begin by dividing the emergency ambulance service system into three component subsystems: Communications, Medical Services and Transportation.* The focus of this document is primarily on the transportation component, and so, after some discussion of the other two, we proceed to a careful examination of the elements of the Transportation system. This involves a consideration of the nature of the demand for emergency transportation, the nature of the response, the consequences of constraints that result in delays in response, and

* This division is due to Savas [see Ref. 21] .

the alternatives to the current methods of responding to emergency demand. By the end of this chapter the ground will have been laid for the development of models in Chapters 4 and 5 to aid in the allocation of ambulances in cities and towns.

3.2 The Emergency Ambulance Service System

Emergency ambulance service is a part, although seriously neglected, of the overall Medical System. Its purpose is to deliver the victim of an emergency, in the best possible state given the circumstances of the emergency, to an appropriate medical care facility. The statement of this objective leaves a lot of room for interpretation in specific cases. It usually requires the rapid transport of trained attendants and equipment to the scene of the emergency. It does not usually require very rapid transport of the patient to the nearest hospital. If the patient's condition has been stabilized at the scene, the advantage of delivering him at a hospital with facilities appropriate to the treatment of his injury or complaint will generally outweigh the disutility of a longer time spent in the ambulance.

We shall find it convenient for the purpose of discussion to divide the ambulance system into the three components mentioned in the Introduction: Communications, Medical Services and Transportation.[21]

3.2.1 Communications

The Communications sub-system includes the means and methods by which information about the existence, location and nature of the

emergency is transmitted through the emergency ambulance system. Figure 3.1 is a very simple representation of the system. At the center is the Dispatcher who is first notified of the existence of the emergency by a citizen observer (usually by telephone), by the police (by radio), or by a telephone operator. In many large cities the dispatching of emergency ambulances is done by the police dispatcher, whether or not the police actually provide the ambulance service. Where the Fire Department provides the ambulance service, the established dispatching facilities of the Department are used. As in the case where a private company provides the service, the Fire Department usually has a direct radio link to Police Headquarters, where a great many emergency ambulance calls are received.

The Dispatcher may attempt an assessment of the seriousness of the call, but in general concentrates on getting details on the nature and location of the emergency from the caller. Contact is then established with an ambulance, either verbally if the vehicle and the Dispatcher are at the same location, otherwise by radio. In the event that no ambulances are available, the Dispatcher might contact a neighboring ambulance service, either by radio or by telephone, and request an ambulance from that service.

Once the ambulance has been directed to the emergency, the Dispatcher contacts the hospital to which the patient is to be taken in order to allow preparation for the arrival. At present this is usually done by telephone, either through normal telephone channels or via one of the "hot lines" (direct telephone lines)

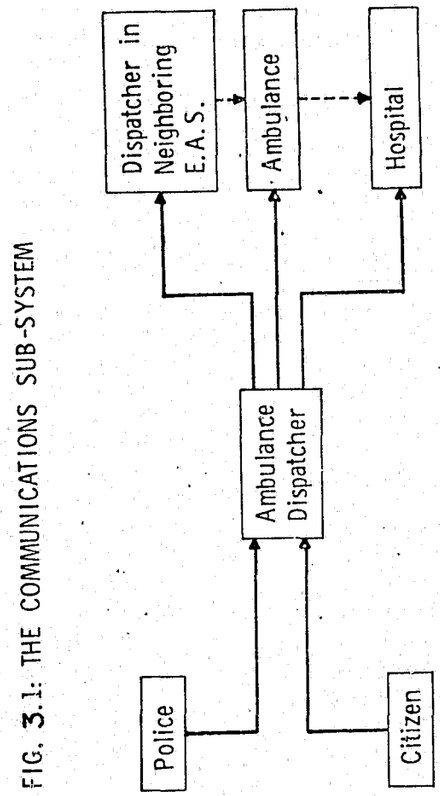


FIG. 3.1: THE COMMUNICATIONS SUB-SYSTEM

that link some Dispatchers with the hospital emergency rooms in their area. The Dispatcher may call the hospital again if radio contact with the ambulance gives rise to additional important information about the emergency. It is conceivable that in the future there will be communication between the ambulance and local hospitals (broken line in Figure 3.1) that will allow doctors at a hospital to monitor the patient's condition and advise ambulance attendants on appropriate treatment.

In those urban emergency ambulance services with which the author is familiar, most of the elementary functions described above are carried out, although notification of the hospital of a patient's arrival is sometimes neglected. Some of the early "scandals" associated with small rural services revolved around the fact that the various parts of the system could not communicate in the manner we have described, and as a result patients suffered long delays before ambulances arrived, or were delivered to hospitals that were unprepared or incapable of receiving them.

The most serious problem associated with the communications function is that of trying to judge the nature of a call for emergency service, and to make some assessment of its priority with respect to calls that might arrive in the near future. Because the responsibilities associated with misjudging the seriousness of a call and delaying ambulance dispatch are so great, the Dispatcher makes little, if any, attempt to distinguish among calls. Most calls are treated as being of equal priority, which implies that any delays are imposed

with equal chance on serious calls (where they may have dire consequences) and on non-serious calls (where there may be no effect at all).

3.2.2 Medical Services

The elements in the Medical Services sub-system are the most intangible of all. The quality of these services provided by the emergency ambulance system is the criterion in terms of which this component of the whole system is judged. Quality of service is an ultimately subjective measure, being a conglomerate of the training, experience and skill of the attendants, the equipment carried on the ambulance, the design of the vehicle, and the degree of access to appropriate expert medical care.

One of the most obvious failings of emergency ambulance services in the past has been that attendants were often unsuited for the job and very poorly trained.[1,15,22] The prominence given to emergency ambulance service recently has had a number of effects:

(i) Pressure has increased to make the position of ambulance attendant a para-medical one, giving the attendants increased responsibilities and opportunities for making a career in the service. This will mean raising salaries at least to the levels received by policemen and firemen, and will probably force the service to be taken over or at least heavily subsidized by the community.

(ii) Under the auspices of the Department of Transportation training programs for ambulance attendants are being improved, and publicity is being given to existing emergency ambulance services that do operate particularly well so that they may be imitated.[2]

(iii) The design of the ambulance has been improved, with functionality rather than appearance determining the shape of the vehicle. A much larger interior space allows the relatively unimpaired administration of first-aid inside the ambulance, the storage of more equipment, and the transport of more than one patient at one time. The last is particularly important because automobile accidents (responsible for 25% of the patients transported in one service observed by the author) often involve more than one person.

Technological developments in medical telemetry may soon allow doctors, monitoring the patient's condition from a distant hospital, to supervise more sophisticated treatment at the scene than can be provided by the attendants alone. These developments may be the only hope of reducing the number of deaths from cardiac failure (by far the biggest cause of deaths among urban emergency victims) where drastic medical action has to be taken very soon after the onset of the condition if the patient is to survive.

3.2.3 Transportation

The important elements in the Transportation sub-system are the

ambulances, their number and location, and the location of the hospitals in the community. Other elements are the area of the region served, the availability of good roads, traffic conditions and other impediments to travel (rivers, mountains, etc.).

We discussed improvements in the design of ambulances and mentioned possible technological innovations in the previous section. One current innovation in emergency transportation has been the introduction of helicopter ambulances. While their use in an urban environment appears to be limited to highway rescue operations, in sparsely populated areas and near large water surfaces they may be extremely effective in an ambulance role. Because of their high costs (capital outlay is approximately $100,000 for one helicopter, and operating expenses are at least double those for a ground ambulance) they have not been extensively incorporated in large services. The city of Chicago has one of the most active helicopter ambulance services (399 calls in 1967), but 75% of the helicopter emergency trips undertaken in Chicago in 1967 were connected with boat mishaps in Lake Michigan.[2] We shall therefore not devote very much attention to their use as ambulances in our consideration of urban emergency services.

There remain the problems of how many (ground) ambulances should be located in which part of the region being served, and what, if any, their spatial relationship should be to local hospitals. We have seen that one of the important reasons why some private operators provided poor service in the past was financial. About 80% of the costs of operating an emergency ambulance service are

connected with personnel salaries and benefits which depend on the number of ambulances allocated to the daily work shifts. It therefore seems reasonable to devote the energies of this document to the exploration of improved ambulance allocation procedures that would reduce the delays suffered by patients and the costs encountered by operators.

3.3 Analysis of Emergency Ambulance Transportation

The time-sequence of events involving the emergency ambulance service that are initiated by an emergency is illustrated in Figure 3.2. Following the occurrence of an emergency there is a delay until it is detected and contact is made with the Ambulance Dispatcher. This delay may be a matter of one or two minutes in a busy metropolitan street, or possibly hours on an isolated highway. If an ambulance is available it is dispatched almost immediately. Otherwise there is a time lag until a previously busy ambulance reports that it is free (Dispatch Delay). Once it has been dispatched the ambulance consumes time travelling to the scene of the emergency (Travel Delay). This time is a function of the relative locations of incident and ambulance, the local street patterns and traffic conditions. At the scene of the emergency the attendants render first-aid in order to stabilize the patient's condition prior to transporting him to an emergency ward.

The primary objective of the Transportation sub-system is to respond as rapidly as possible to the demand for emergency service. Doctors have stressed the need for the arrival of medical aid as soon after the

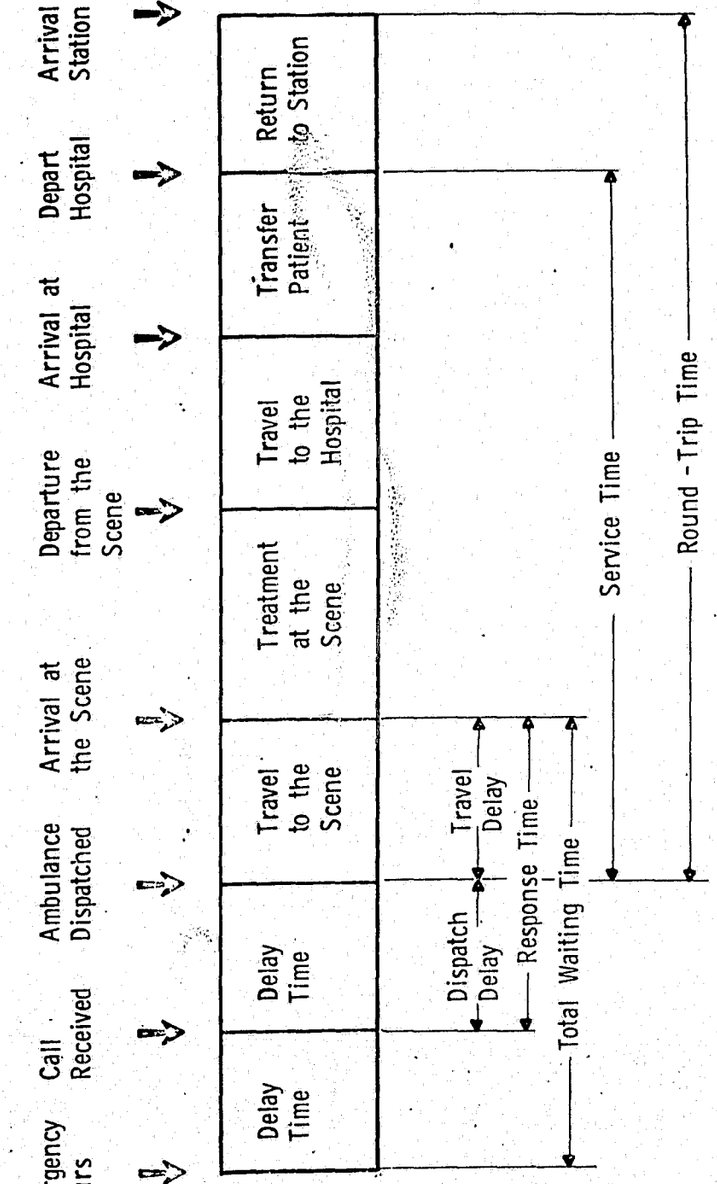


FIG. 3.2. SEQUENCE OF EVENTS FOLLOWING AN EMERGENCY

occurrence of the emergency as possible.[14] In this study we shall be primarily concerned with allocating ambulances in such a way as to keep the Response Time (see Figure 3.2) down to a reasonable level. In Chapter 4 we develop some simple measures relating the number of ambulances in the emergency system and their location to the Response Time.

A secondary objective of the Transportation sub-system is the reduction of the Service Time - the total time that an ambulance is occupied with one call (see Figure 3.2). Reducing the Service Time increases the probability that an ambulance is available to respond immediately to an emergency call. This reduction can be achieved by reducing unnecessary delays in ambulance response and in the time spent travelling from the scene to a hospital. Unnecessary delays in ambulance response are the result of too few vehicles participating in the service and/or their being poorly located relative to the incidence of emergencies. This problem will be examined in the succeeding chapters. Unnecessary delays in transporting the patient are the result of travelling to hospitals located far away from the incident while viable alternatives exist close at hand. Appendix B is addressed to this issue, and describes a method for selecting the hospital nearest to the incident.

We shall now focus on the Response Time, specifically on the Dispatch and Travel Delays, outlining the problems that are encountered in developing stochastic models. We begin with a general discussion of the nature of the demand for emergency ambulance service

3.3.1 The Arrival of Emergency Calls at the Dispatcher

From the author's observations in and around the city of

Boston, it seems reasonable to assume that emergencies are generated randomly, i.e., the time and location of any emergency are independent of the times and locations of all other emergencies. There are obvious exceptions to this general rule, for example, natural disasters (hurricanes, floods, earthquakes), airplane crashes, large structural fires and multiple highway accidents, but these are rare events, and so we shall assume the random (Poisson) arrival of calls in our modelling efforts.

Randomness does not imply uniformity, and, in concert with the fire and police services, a problem for emergency ambulance operators accommodating the almost continuous change (both spatial and temporal) in the rate of call generation. Figure 3.3 illustrates the temporal changes in the demand encountered by the emergency ambulance service in Cambridge, Mass. during a 24-hour period. In addition there are variations by day of the week, and also seasonal fluctuations. Superimposed on the temporal variability are the differences in the rates at which calls are generated from different parts of the region served. These differences are due to population densities, socio-economic factors and the presence or absence of busy highways. These differences are not clearly understood, although studies carried out by the author suggest that low-income neighborhoods generate more emergency calls per capita than high-income neighborhoods. This is apparently due to the fact that the poor have little or no access to private physicians and therefore have to be taken to a hospital for medical treatment.

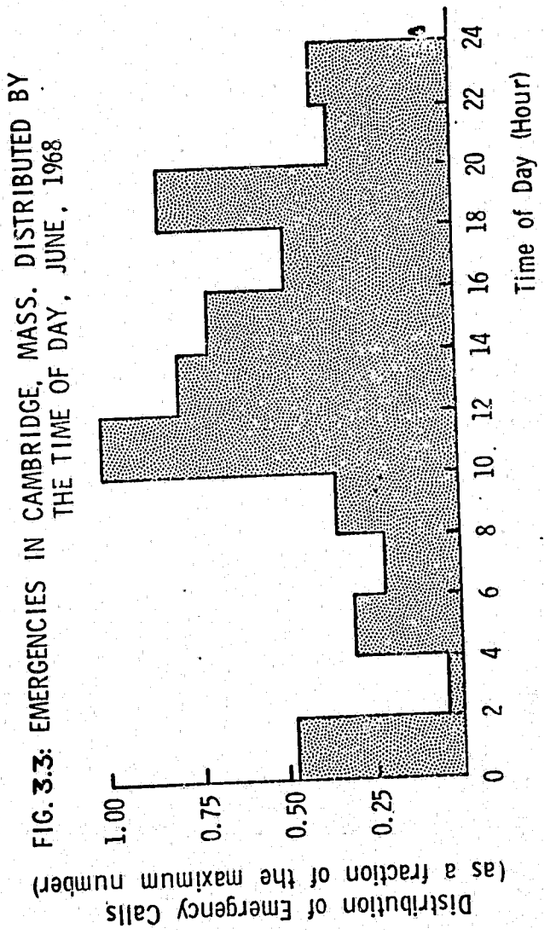


FIG. 3.3: EMERGENCIES IN CAMBRIDGE, MASS. DISTRIBUTED BY THE TIME OF DAY, JUNE, 1968

Sometimes the changes in the temporal and spatial demand rates may not be independent. For example, the demand for emergency service existing in the business district of a large city during the day becomes almost negligible when the office workers leave the area at night. There may then be an increase in the occurrence of emergencies at night, especially over weekends, in those parts of the city devoted to entertainment.

The realities of staffing an emergency service do not permit the continuous changing of the response force to meet a continuously changing demand rate. Except in very small towns served by volunteers called from their homes, we can reasonably assume that the staffing of an emergency ambulance service is in terms of two or three daily shifts. In making our allocation of ambulances and attendants to each shift we shall assume that the day can be divided into n equal shifts so that the weighted sum of the variances of the demand rate in each shift is minimized for any n. Ideally we should like n to be large. In practice n will be constrained to be 2, 3 or perhaps 4. We shall then use the mean demand rate in each shift to make allocations to those shifts.

3.3.2 The Type of Emergency Call

Not all calls for emergency ambulance service are equally urgent. In fact, with the disappearance of the general practitioner who used to treat minor emergencies (cuts, sprains, dog bites, etc.) in the home, there has been a recent increase in the incidence of

non-urgent emergencies serviced by the emergency ambulance service.[23] Distinguishing between urgent and non-urgent emergencies would allow greater system flexibility resulting in lower costs and better service to the victims of serious emergencies. Non-urgent emergencies might be transported in vehicles that would provide comfort and support without satisfying all the specifications of Appendix C. These vehicles could be staffed by personnel who had not acquired all the skills of experienced ambulance attendants, but were trained to deal with the victims of non-urgent emergencies. In Cambridge, Mass. fully 50% of all those transported by the emergency ambulance service are the victims of non-urgent emergencies. They are transported by the police at a significant saving in cost to the city.

If the Dispatcher could assess a priority on an incoming call for service, it would be possible to adjust the type of response accordingly. The Dispatcher might even delay the response to a low-priority call during a busy period even though an ambulance were available. This would leave the vehicle free to answer a higher priority call arriving subsequently. In practice, making an assessment on the urgency of an emergency call over the telephone is extremely difficult, and carries with it serious legal and ethical responsibilities. Therefore there is little attempt to assess priority, and in our modelling we shall assume that in general the Dispatcher does not have this option.

3.3.3 The Waiting Time

Between the occurrence of an emergency and the arrival of an

ambulance at the scene there are three components of what we have called the Waiting Time in Figure 3.2. It is generally recognized in the medical profession that this delay is a critical period in the passage of a patient through the emergency medical system, and should be kept as small as possible.[14]

The three components of delay are:

(i) The delay until notification of the emergency ambulance service of the incident. (Notification Delay)

(ii) The delay until an ambulance becomes available for dispatch to the incident. (Dispatch Delay)

(iii) The delay between the dispatch of the ambulance and its arrival at the incident. (Travel Delay)

We shall focus our attention on the last two components of the Waiting Time in our modelling efforts because they appear to offer the greatest hope of reduction. We can discuss the components of the Waiting Time briefly as follows:

(i) Notification Delay: In most urban areas this delay is usually small because of the proximity of other people. On isolated highways, however, notification presents an acute problem as is indicated by the high proportion of rural highway accidents that result in death. Current attempts to reduce the Notification Delay involve the installation of telephones along highways, experiments with alarm devices triggered by impact of a highway guard-rail, and increased aerial and ground patrol by highway police. With the introduction of the

universal emergency telephone number, 911 (operative in New York City at this time and expected in other cities around 1972) notification of the emergency ambulance service will be expedited. In the past, because of the low priority given to ambulance service, and because of the great variety of purveyors, members of the public were often at a loss in deciding how to go about contacting the emergency ambulance service.

(ii) The Dispatch Delay: Any emergency ambulance service consists of a number of ambulances which respond to randomly arriving calls for service, and are exclusively occupied with each call until the admission of a patient to an emergency ward. Now it is quite feasible that for some period of time all the ambulances might be busy. Every call arriving in this period will be delayed until an ambulance becomes free to respond to it. We are assuming that there are no alternative sources of ambulances.

The number of calls that wait and the length of time that each waits is a function of the rate at which calls arrive, of the time taken to serve a call, and of the number of ambulances in the service. Conceivably, the Dispatch Delay might also be affected by the order in which the calls are answered. The Dispatcher might assign the most urgent from among a queue of delayed calls to the first available ambulance regardless of its position in the queue. Another possible strategy is to assign to an ambulance that call from

the queue of delayed calls which is closest to the vehicle. While these are both feasible strategies, they are not generally used. Instead a first-come, first-served discipline prevails.

(iii) The Travel Delay: While the Dispatch Delay is encountered by some patients (unless the ambulance service has so few ambulances that there is always a queue of patients implying that <u>all</u> patients experience a Dispatch Delay), the Travel Delay is experienced by every patient. The Travel Delay is a function of the relative locations of ambulance and incident, the layout of streets and traffic conditions in the area. As traffic congestion in the large cities gets worse the Travel Delay component of the Waiting Time increases. This has led to the consideration of helicopters as ambulances, especially in connection with highway accidents. In New York City some doctors are travelling by foot and by subway to get to the victims of cardiac failure quickly. In the near future we may see ambulance attendants on motor-scooters beating the traffic and the ambulance to an emergency victim.

Unfortunately, very little has been done in most emergency services to relate the location of the ambulances to the incidence of emergency calls. This has resulted in unnecessarily long Travel Delays that could be reduced by a simple relocation of the vehicles.

3.3.4 The Service Time

The last element of the Transportation sub-system that we shall consider is the Service Time (see Figure 3.2). The Service Time is the sum of the time taken by the ambulance and its crew after dispatch to reach the patient, the time to administer first-aid at the scene, and the time to transport the patient to a hospital. Since we assume that call arrivals are spatially independent, it seems reasonable to assume that the Service Times are independent.

We undertake a fairly detailed analysis of the Service Time in Appendix A, developing a description of the Service Time distribution. From this work the Service Time is certainly not distributed as the negative exponential. Nevertheless, to facilitate our initial modelling efforts in Chapter 4, we shall assume that the Service Time is distributed as the negative exponential. We shall be able to discard the assumption in Chapter 5.

3.4 Chapter Conclusion

Separating the Emergency Ambulance Service System into its three component sub-systems: Communications, Medical Services and Transportation, has allowed us to consider briefly the important elements of each. It appears that the Transportation sub-system has elements (amenable to analysis) which offer the best opportunity of improving system performance by explicitly relating allocation to demand and reducing costs through the more efficient use of resources.

The ultimate measure of the performance of the emergency ambulance service is the condition of the patient in the hospital emergency ward. Since this is a somewhat intractable measure, we need to develop surrogate measures that give some indication of the patient's condition. One of these measures is the Waiting Time (Figure 3.2). A long Waiting Time in general has an adverse effect on the emergency victim's ultimate condition, and therefore constitutes a measure of the Transportation sub-system's performance. Ignoring Notification Delay allows us to use the Response Time (Figure 3.2) as a measure.

We therefore proceed to Chapter 4 with the picture of the sequence of events following an emergency afforded by Figure 3.2, the assumption of random call arrivals and exponential service times. In Chapter 4 we shall analyze the components of the Response Time and their relation to system configuration. A result of the analysis will be a procedure whereby ambulance service administrators may rationally allocate vehicles in response to demand, taking account of the size of the area being served.

CHAPTER 4: MODELS FOR AMBULANCE ALLOCATION

4.1 Introduction

In Chapter 3 we identified the Response Time of an ambulance to be one possible criterion by which to measure system performance. Consisting of two components, the Dispatch and the Travel Delays, the Response Time is a function of the number of ambulances in the area served, their location, the spatial and temporal distribution of emergency calls and traffic and road conditions in the area. In this chapter we consider two very simple models which allow us to approximate the number of ambulances (and their location) required to provide either (i) an acceptably low response time or (ii) the lowest attainable response time given municipal budget constraints. We shall assume throughout the chapter that emergency ambulance service is provided either free of charge by the community, or at a fixed rate per patient with bad debts absorbed by the community.

We begin by modelling the Dispatch Delay in Section 4.2, invoking results from the theory of queues. From the model we develop measures of ambulance availability, average dispatch delay, and ambulance utilization in terms of the demand for emergency service and the number of ambulances allocated to the area. In addition the model provides useful insights into the problems encountered by small private operators trying to provide an emergency ambulance service.

In Section 4.3 we make assumptions about the layout of streets in a city and model the time taken by an ambulance to travel to the scene of

an emergency. It is a simple matter to show that dispersing ambulances around an area reduces travel time compared to a single location. Using a result derived for allocating police patrol cars we can find an upper bound on the time for the nearest available ambulance to reach a call.

Combining the results of the two models allows the calculation of an expected response time, and the allocations of ambulances around the city such that city-wide response time is minimized. A dynamic programming method of making the allocation is demonstrated at the end of the chapter by means of a simple example.

Because of the restrictive assumptions necessary in order to apply analytical models to real-world situations (regularity of street pattern, constant average travel speed, approximately constant average call rate, etc.) the results of these models cannot be taken as final. They are, instead, guides to planning an emergency service, indicative of the approximate requirements of a reasonable service. Once the service is in operation, ambulance allocations for a particular workshift may have to be changed slightly because of other considerations. However, by taking into account as many of the variables as possible, we can make estimates of ambulance requirements that will allow budget prediction and planning in response to changes in demand and the cost of service.

4.2 The Dispatch Delay

Incorporating the assumptions made about the generation of calls (i.e. Poisson) and the probability distribution of the service time (i.e. negative exponential), we shall now proceed to develop a model that will

allow the estimation of the probability that an emergency patient encounters a delay before an ambulance can be dispatched, and the expected delay time. We assume that N ambulances are assigned to a region during a working shift to respond to randomly arriving calls. If an ambulance is available when a call arrives it is dispatched immediately and travels to the scene. If all the ambulances are occupied on other calls, then the patient has to wait until an ambulance becomes free to respond. If a number of calls for emergency service arrive while the system is saturated, they form a queue and are answered according to a first-come, first-served rule. An ambulance is unavailable when it is travelling to the scene of an emergency, treating a patient at the scene, or transferring the patient to a hospital. We shall not consider the pre-emption of an ambulance en route to one emergency by another, more serious, call since these appear to constitute a very small fraction of all calls.[24] These assumptions are summarized in Table 4.1 for the reader's convenience.

4.2.1 The Probability of a Delay

The assumptions made above define the extensively studied N-server queuing model. With calls arriving in a Poisson manner at a mean rate λ calls per hour, each being served in a time drawn independently from the negative exponential distribution with mean $\frac{1}{\mu}$ hours, we may quote the following result from a standard text on queuing theory*.

---

*Reference No. 20, page 343.

Table 4.1: Summary of Assumptions in Ambulance Dispatch Delay Model

1. Calls for emergency service arrive in a Poisson manner at mean rate λ calls per hour during a given working period.
2. A total of N ambulances are assigned to respond to these calls.
3. The total time to service an emergency call is independently drawn from the negative exponential distribution with mean $1/\mu$ hours.
4. A call is delayed when all the N ambulances are unavailable for dispatch.
5. Delayed calls form a queue and are answered as ambulances become available in a first-come, first-served manner.
6. There are no external sources of ambulances.
7. Once assigned to a call an ambulance does not desert that call for another.

When the service is in the steady-state, the probability that an emergency call experiences a dispatch delay, Q_N, is a function of λ, μ and N as follows*:

$$Q_N = \frac{\rho^N}{N!} \left(\frac{N}{N-\rho}\right) \left[\sum_{i=0}^{N} \frac{\rho^i}{i!} + \left(\frac{\rho}{N-\rho}\right) \frac{\rho^N}{N!}\right]^{-1} \qquad (4.1)$$

For convenience we have used ρ, defined as the "demand intensity," where $\rho = \frac{\lambda}{\mu}$. Equation 4.1 is the well-known Erlang Delay Formula used in telephone traffic theory.

Q_N is plotted in Figure 4.1 against the demand intensity ρ with N, the number of ambulances as a parameter. Each combination of ρ and N determines a particular value of Q_N. We should like to keep the probability of a dispatch delay as low as possible. From Figure 4.1 we can estimate the value of Q_N for any value of ρ in the range 0 to 7. We shall define Q_N to be the "level of service" at the specific value of ρ encountered by our service. For example, if the mean call rate during one shift is such that the demand intensity ρ is equal to 3, then an ambulance service with 6 vehicles will provide a 10% level of service (see Figure 4.1). This means that the steady state probability of a call encountering a delay is less than or equal to 10%. Alternatively, an <u>average</u> of 10% of all calls arriving during the shift will encounter a dispatch delay.

In Figure 4.2, which is easily derived from Figure 4.1, we indicate the number of ambulances needed to provide respectively (at

*A list of the symbols used in this chapter is included as Table 4.4 on page 91.

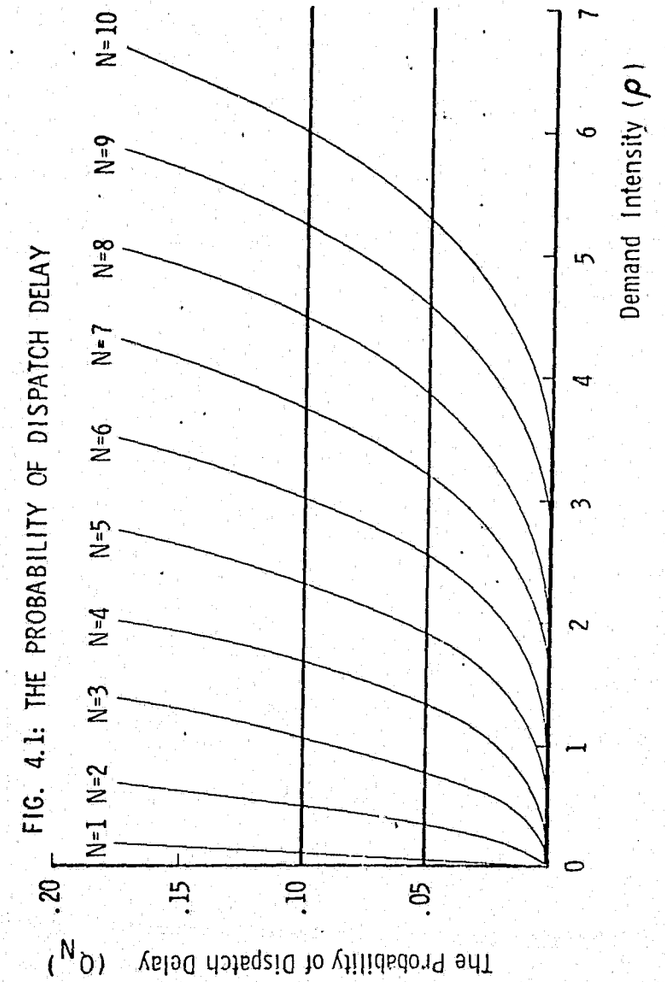


FIG. 4.1: THE PROBABILITY OF DISPATCH DELAY

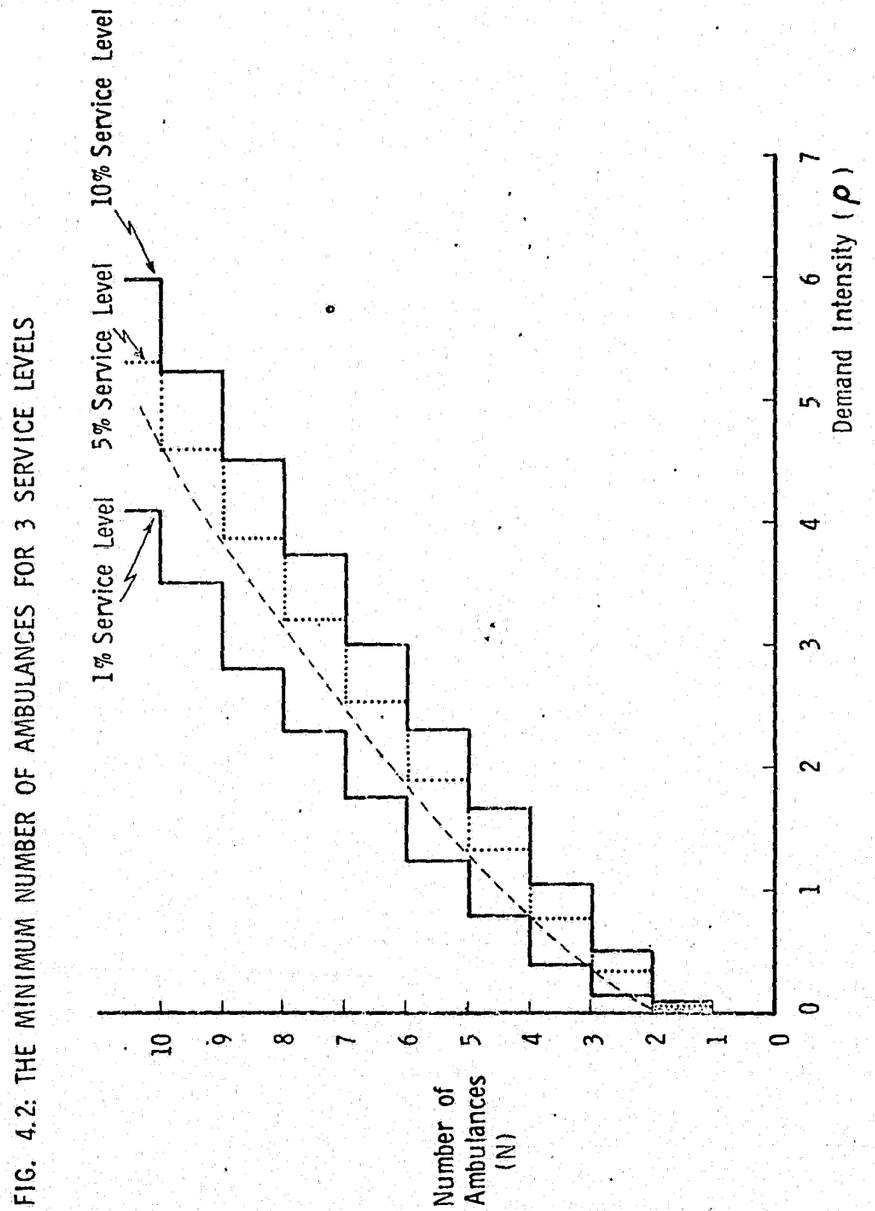


FIG. 4.2: THE MINIMUM NUMBER OF AMBULANCES FOR 3 SERVICE LEVELS

least) a 1%, 5% and a 10% level of service. With a mean call rate such that $\rho=4$, we should require 8 vehicles at the 10% level, 9 ambulances at the 5% level and 10 ambulances at the 1% level of service. According to the desired level of service, ambulances may be allocated to each working shift on the basis of the anticipated demand intensity, ρ for that shift.

4.2.2 The Expected Dispatch Delay

From the same queuing model that gave us equation 4.1 we can find the conditional expected dispatch delay, i.e., the expected dispatch delay encountered by that group of patients who experience a non-zero delay. From Reference No. 20, page 345, we see that the conditional expected dispatch delay is

$$E[D_c] = \frac{1/\mu}{N-\rho} \tag{4.2}$$

where all the quantities are as defined previously.

The unconditional dispatch delay is simply given by the product of 4.2 and the probability of a dispatch delay, Q_N. Therefore, the unconditional expected dispatch delay is

$$E[D] = \frac{Q_N}{\mu(N-\rho)} \tag{4.3}$$

where Q_N is defined by equation 4.1.

We observe that the conditional dispatch delay, which is a better measure of the actual delay encountered by patients than the unconditional delay, is a function of the mean service time, and

depends inversely on the number of ambulances. We plot $E[D_c]$ in Figure 4.3 against demand intensity ρ with N as a parameter. On this graph we superimpose the conditions representative of the 1%, 5% and 10% service levels. We can make three observations from the graph that suggest that the "level of service" concept must be used with great care:

(i) At the same level of service $E[D_c]$ is lower for services with a high demand intensity, suggesting that larger areas with high demand are better than smaller ones.

(ii) For very small services (1, 2 or 3 ambulances) $E[D_c]$ is extremely large even at the 1% level of service, suggesting an inadequacy of that measure in this case.

(iii) For large services, the introduction of another ambulance at the same demand intensity does little to reduce the dispatch delay for those who encounter it. (Although the fraction of all patients encountering such a delay falls.)

4.2.3 The Utilization of the Ambulance Service

We shall define the Utilization U of the ambulance service as the ratio of the total number of ambulance hours spent servicing calls (transporting and treating patients) in a working period to the total number of ambulance-hours available in that period. In a period of length T, an N-ambulance service has NT ambulance-hours available. The number of ambulance-hours actually expended servicing emergency calls can be represented by H_w, where

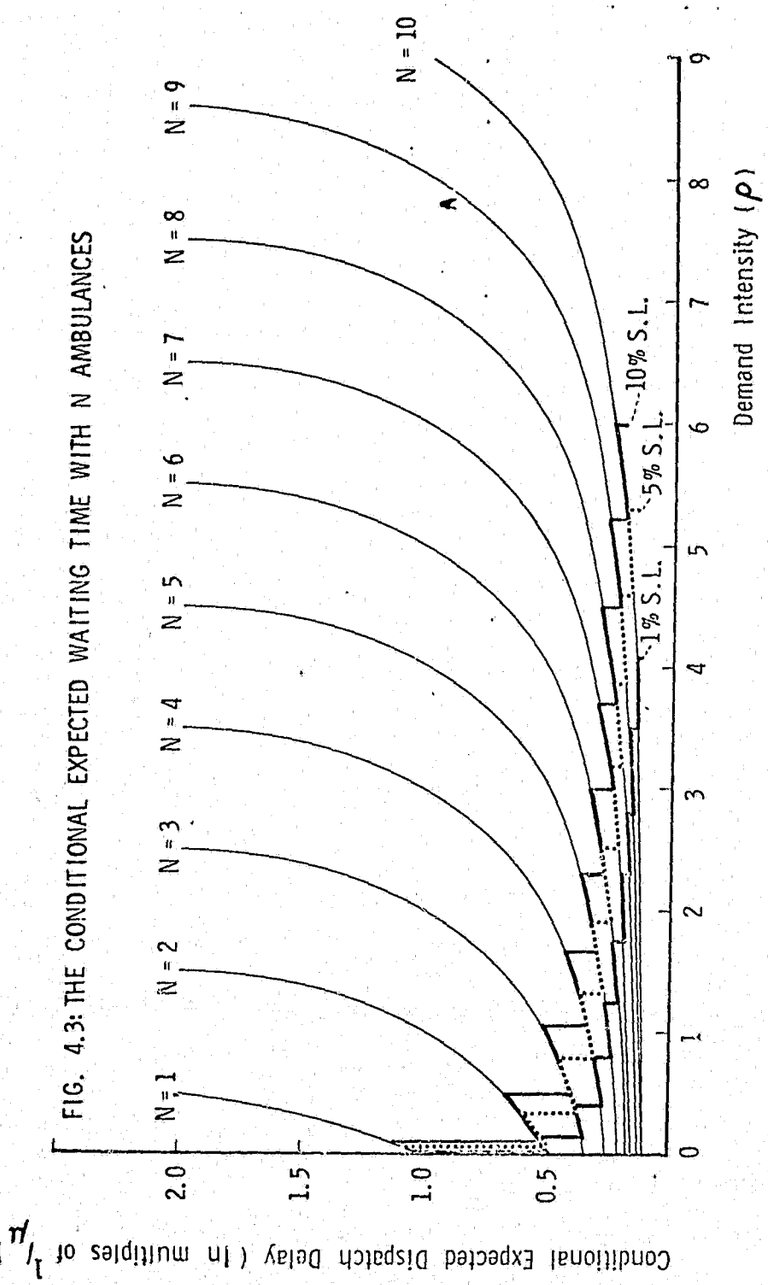


FIG. 4.3: THE CONDITIONAL EXPECTED WAITING TIME WITH N AMBULANCES

$$H_w = [\sum_{i=0}^{N-1} iP_i + N \sum_{i=N}^{\infty} P_i]T$$

where P_i is the steady-state probability that there are i patients being served and waiting to be served (i=0,1,2,...). From the same theory as gave us equation 4.1, we can show that

$$H_w = [(1-Q_{N-1})\rho + N\, Q_N]T$$

Therefore $$U = (1-Q_{N-1})\frac{\rho}{N} + Q_N \qquad (4.3)$$

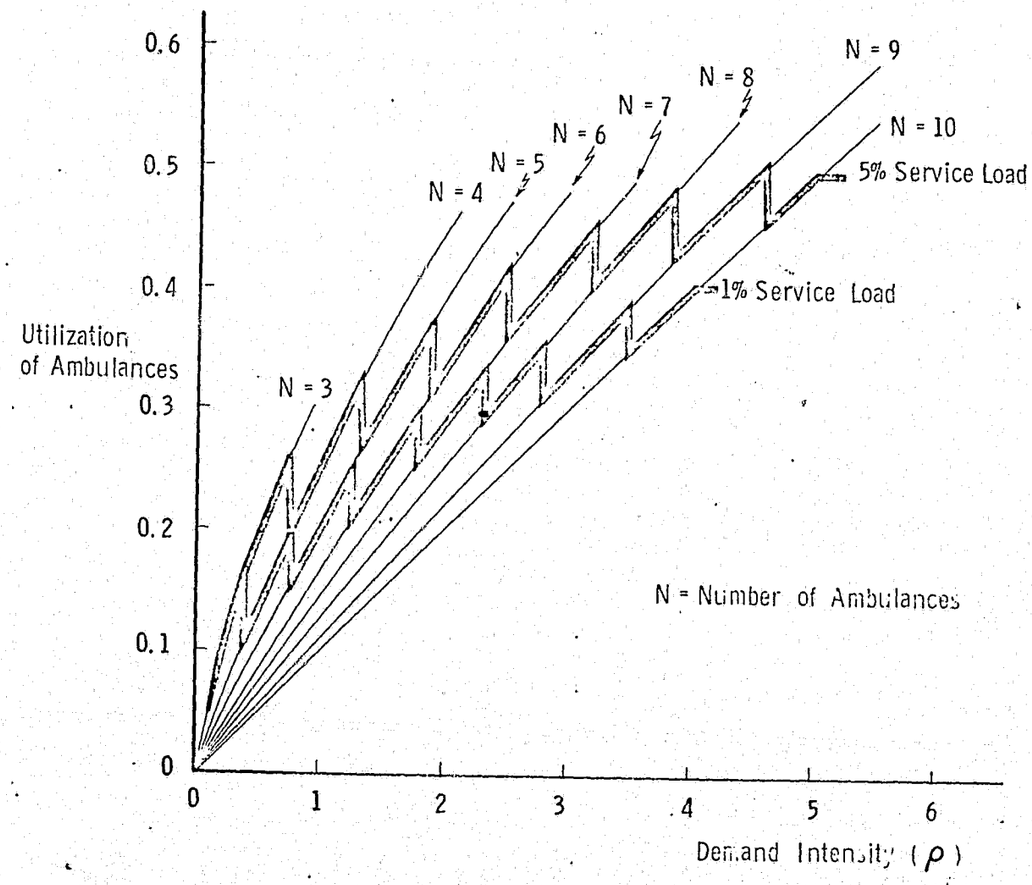
For small Q_N, Q_{N-1} (i.e., when operating at a high service level)

$$U \approx \frac{\rho}{N} \qquad (4.4)$$

Equation 4.4 implies that an estimate of the Utilization can be made from Figure 4.2, for say Q_N equal to 1% and 5%, simply by dividing the value of ρ by the number of ambulances at a particular service level.

In Figure 4.4 we plot equation 4.3 as a function of ρ, with N as a parameter and the conditions representing the 1% and 5% service levels superimposed on the main graph. From this graph an obvious concomitant to a high level of service is an extremely low utilization, and therefore a relatively high operating cost per patient treated. This is particularly pronounced where the demand intensity is low, and helps to explain why morticians, garages and taxi-cab fleets have provided emergency ambulance service in small towns in the past. Whereas this is unfortunate, it is not unusual in other emergency-type services where calls arrive randomly in time. The utilization of fire departments is at a similarly low level. Police

FIG. 4.4: UTILIZATION OF AMBULANCES



patrol cars spend a considerable fraction of the working period in "preventive patrol," a function that fills in time between answering calls for police service.

Where, because of the low utilization, the cost per patient of an acceptable emergency service has become prohibitively high, the only alternative would appear to be the partial or complete subsidization of the service by the community. Coupled with an intensive search for ways to reduce costs (which we shall investigate in later chapters) and some Federal subsidization for the provision of emergency services to highway accident victims, community sharing of responsibility for emergency ambulance service is becoming more and more common.[2,5,8,22]

4.2.4 The Utility of the Dispatch Delay Model

The Dispatch Delay Model developed in this section is useful first as an aid to the administrator of an ambulance service, and second as a source of insight into the operation of the emergency ambulance transportation system. An administrator armed with Figure 4.2 and aware of the behavior of the expected dispatch delay (4.2.2) can make a rough estimate of the number of ambulances needed to provide an acceptable level of service in any working shift. Combining this model with the model of the travel time in the next section produces a more comprehensive measure of the system's performance for the allocation of ambulances. One delicate problem is the definition of an acceptable level of service. It is unlikely

that the distinction between a 1% and a 5% service level is meaningful to anyone. On the other hand the distinction between a 5% and a 25% service level is much clearer, and does allow the allocation of available ambulances and attendants to those shifts in which they are needed most.

The insights into system operation provided by the model may almost be more important than anything else in this instance. We have noted the low utilization of small services endeavoring to provide a high level of service. In addition, the conditional expected dispatch delay for small services ($\rho < 1$) is relatively high, even for very high service levels. Finally it appears that to improve the system's operation at the margin is a costly proposition. Figure 4.2 suggests that in order to improve a service from the 5% to the 1% level requires an increase of about 20% in the number of ambulances. If, as we have suggested, 80% of the cost of the emergency ambulance service is incurred in manning ambulances, then improving service to 4% of the patients requires a 16% rise in costs. While this is not unexpected, it does serve as a spur in the search for an alternative method for responding to calls delayed by system saturation.

4.3 The Travel Delay

The second component of the Response Time in Figure 3.2 is the Travel Delay, the time between the dispatch of an ambulance and its arrival at the scene of an incident. This delay is a function of the

distance between the ambulance at the moment of dispatch and the incident, and of the speed at which the vehicle is able to travel. That speed is itself a function of traffic conditions which will vary considerably during the day in most urban areas.

In order to come to grips with the time spent in travel we shall make some simplifying assumptions that will allow us to explore the consequences of different ambulance location policies. Currently the location of ambulances varies according to the agency providing the service. Police vehicles, for example, patrol the streets continuously, ambulances administered by Fire Departments are housed in fire-stations around the city, private operations usually have one central location, and ambulances associated with hospitals are usually based at one of these institutions.

We shall begin with a very simple descriptor of the distance between the ambulance and the incident, and then examine the expected travel time that results from different locations of the ambulances. Ideas and results derived in Appendix A are used in this section and the reader might want to browse through that portion of the document before proceeding further.

4.3.1 The Travel Distance

We shall assume that our ambulance service is to be located in a city in which the streets are laid out in a rectangular grid pattern, i.e., streets are mutually perpendicular, say running North-South and East-West. All the streets are assumed to be two-way and the distance between any two points is simply the sum of the East-

West and the North-South distances. By establishing the origin of a Cartesian coordinate system in the bottom left-hand corner of our hypothetical city we can represent any point therein by an (x,y) pair. (See Figure 4.5.) The distance between any two points (x_i,y_i) and (x_j,y_j) is given by

$$d = |x_i-x_j| + |y_i-y_j|$$

If v_x and v_y are the variables representing the respective speeds in the x and y directions, then the expected time to travel between (x_i,y_i) and (x_j,y_j) is

$$t_{ij} = E\left[\frac{1}{v_x}|x_i-x_j| + \frac{1}{v_y}|y_i-y_j|\right]$$

We shall assume throughout that $v_x = v_y = v$, so that t_{ij} is simply

$$t_{ij} = E\left[\frac{1}{v}(|x_i-x_j| + |y_i-y_j|)\right]$$

4.3.2 Travel Time as a Function of Ambulance Location

(i) One Ambulance in the City:

If one ambulance has been assigned to the hypothetical city of Figure 4.5, the calculation of the expected Travel Delay is a simple matter. Assume that calls for emergency service are uniformly distributed spatially and occur independently of each other and of the position of the ambulance. From the work done in Section A.2 of Appendix A we may quote the following results:

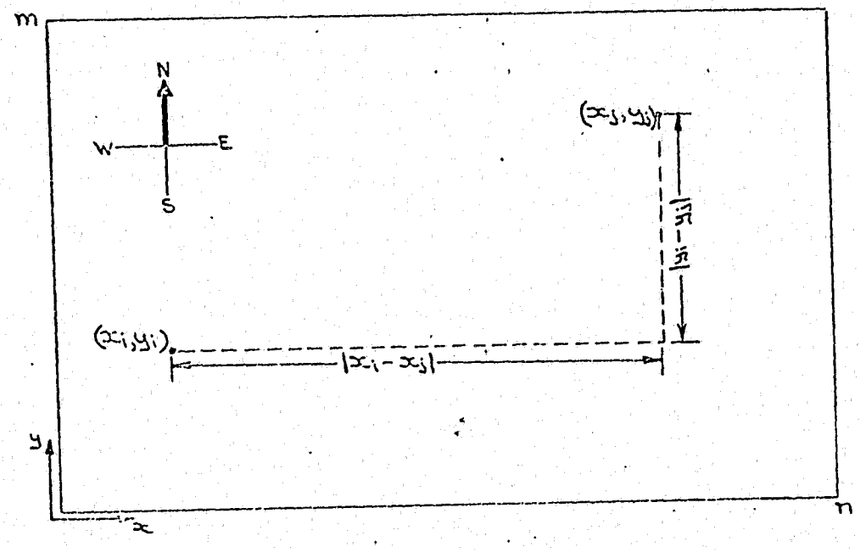


FIGURE 4.5 RECTANGULAR TRAVEL DISTANCE BETWEEN TWO POINTS

(a) If the ambulance is located in the center of the city, the expected Travel Delay is

$$E[t_c] = \frac{1}{4v} \text{ [mtn]} \qquad \text{(from A.6, page 131)}$$

If the region is a square of area A, then we can write

$$E[t_c] = \frac{\sqrt{A}}{2v} \qquad (4.5)$$

This approximation can be shown to be reasonable even when the area is not square[12].

(b) If the ambulance is not located at the center of the city, but instead has its position described by a spatially random distribution, it then follows from A.7, page 132 that the expected Travel Delay is

$$E[t_r] = \frac{1}{3v} \text{ (mtn)}$$

Again if the region is square of Area A, we can write

$$E[t_r] = \frac{2\sqrt{A}}{3v} \qquad (4.6)$$

Thus, if there is only one ambulance in the city, pre-positioning the vehicle at the center leads to a lower Travel Delay than if the vehicle were randomly located. However, if the demand were sufficiently large, there would always exist a queue of waiting calls. If the hospitals are randomly located around the city, the ambulance would respond to each call from some random location and the expected travel time would accordingly be approximated by 4.6.

(ii) N Ambulances in the City

If instead of only one ambulance, N vehicles are allocated to the city the determination of the expected Travel Delay becomes more difficult. There are at least three possible strategies for locating the ambulances:

(a) If all the ambulances are located in the center of the city then, as long as there is at least one ambulance available, the expected Travel Delay is identical to that for the single ambulance

$$\text{i.e.} \quad E[t'_{NC}] = \frac{\sqrt{A}}{2v} \qquad \text{from 4.5}$$

If, at the arrival of an emergency call, every ambulance is busy, then the first vehicle to become available is dispatched, and the expected travel time for that vehicle is identical with that for a randomly located ambulance (on the assumption that the city's hospitals are randomly located)

$$\text{i.e.} \quad E[t''_{NC}] = \frac{2\sqrt{A}}{3v}$$

Therefore, in general, the expected Travel Delay for centrally located ambulances is given by

$$E[t_{NC}] = \frac{\sqrt{A}}{2v}\sum_{i=0}^{N-1} P_i + \frac{2\sqrt{A}}{3v}\sum_{i=N}^{\infty} P_i$$

$$= (1-Q_N)\frac{\sqrt{A}}{2v} + Q_N\frac{2\sqrt{A}}{3v}$$

$$= \frac{\sqrt{A}}{6v}(3+Q_N) \qquad (4.7)$$

where Q_N, the probability of system saturation (or of a dispatch delay) is defined by equation 4.1.

(b) If, instead of being centrally located, the ambulances are dispersed randomly around the city the calculation of the expected Travel Delay becomes more complicated. We assume that, if a number of ambulances are available when a call for service arrives, the nearest vehicle is dispatched to the call. If no ambulances are available, the first vehicle to become free is dispatched. A queue of calls is dealt with on a first-come, first-served basis.

If, at the instant an emergency call is received, there are n ambulances available ($0 < n \leq N$), then we need to find the distribution of the minimum of the n distances between the incident and the randomly located ambulances. Larson[13] has shown that this may be approximated by the Rayleigh* distribution with parameter $2\sqrt{\frac{n}{A}}$. Therefore the probability density function for the minimum travel time t'_{NR}, given n of the N vehicles available ($n > 0$) is

$$f_{t'_{NR}}(t) = \frac{4nv^2t}{A}e^{-nv^2t^2/A} \qquad t \geq 0$$

From which, it follows that

$$E[t'_{NR}] = \frac{1}{4v}\sqrt{\frac{2\pi A}{n}}$$

If no ambulances are available then again the expected travel time is that of a randomly located ambulance.

---

*Rayleigh Probability Density Function: $f_x(x_o) = a^2x_o e^{-a^2x_o^2/2}$ [Ref. 4]. Note that this approximation is somewhat unsatisfactory for small n.

i.e. $E[t''_{NR}] = \frac{2\sqrt{A}}{3v}$

We may combine these as before to get the expected Travel Delay

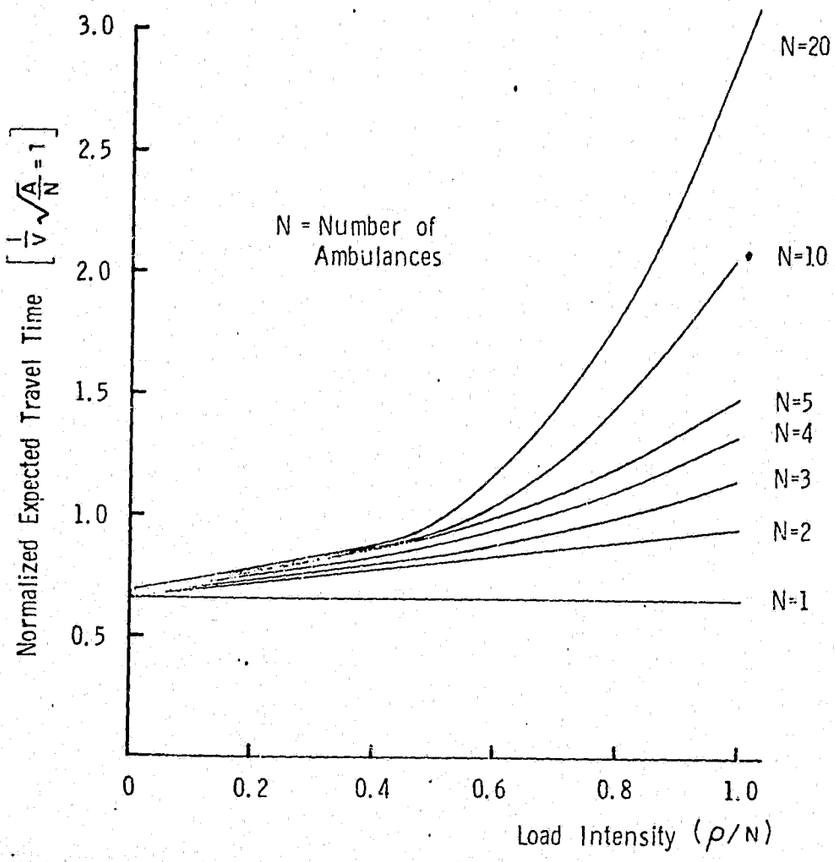
$$E[t_{NR}] = \sum_{i=0}^{N-1} P_i \frac{1}{4v}\sqrt{\frac{2A}{N-i}} + \frac{2\sqrt{A}}{3v}\sum_{i=N}^{\infty} P_i \qquad (4.8)$$

where P_i is the steady-state probability that there are i patients being served and waiting to be served (i=0,1,2...). A term-by-term comparison of 4.7 and 4.8 shows that $E[t_{NR}] \leq E[t_{NC}]$, indicating that a distribution of the ambulances around the city reduces the expected Travel Delay over a central location. Equality of the two delays occurs only when the system operates under conditions approaching continual saturation.

Equation 4.8 is plotted in Figure 4.6 against the load intensity $\frac{\rho}{N}$. (Figure due to Larson, Reference No. 13, page 330.) The expected travel time $E[t'_{NR}]$ has been normalized by expressing it in terms of $\frac{1}{v}\sqrt{\frac{A}{N}}$.

(c) Finally, if, instead of randomly locating the ambulances, we pre-positioned each vehicle regularly throughout the city we could anticipate reducing the expected Travel Delay. Unfortunately the mathematical description of such a location pattern and the interactions among the incidents and the ambulances is extremely difficult. So we shall resort to using equation 4.8 as an upper bound on

FIG. 4.6: NORMALIZED EXPECTED TRAVEL TIME FROM RANDOMLY LOCATED AMBULANCES TO RANDOMLY DISTRIBUTED EMERGENCIES (Due to Larson, Ref 13)



expected Travel Delay, keeping in mind the approximation when using the expression for $E[t_{NR}]$.

4.4 The Response Time

The calculation of the Response Time is now a simple matter of combining the Dispatch and Travel Delays from the previous two sections. We shall carry forward all our assumptions from these sections in deriving the Expected Response Time. Since dispersing ambulances around the region served results in a lower expected Travel Time than the alternative single central location, we shall use equation 4.8 in our calculations.

In a city like that depicted in Figure 4.5 and described in Section 4.3.1, where emergency calls arrive at mean rate λ in a Poisson manner and are uniformly distributed over the space, then, if ambulances are randomly distributed throughout the city, the Expected Response Time is simply the sum of 4.3 and 4.8

i.e., Expected Response Time, $E[R] = E[D] + E[t_{NR}]$

$$\text{i.e., } E[R] = \left[\frac{Q_N}{\mu(N-\rho)}\right] + \left[\sum_{i=0}^{N-1} P_i \frac{1}{4v}\sqrt{\frac{2\pi A}{N-i}} + \frac{2\sqrt{A}}{3v}\sum_{i=N}^{\infty} P_i\right] \qquad (4.9)$$

Now, in very few cities will emergencies be generated at the same rate everywhere. To take account of this variation we shall divide the city into regions such that the mean call rate λ_j in region j is roughly constant across the region in any working shift. We can then allocate ambulances randomly across each region, and assume that the mean service time is equal to $\frac{1}{\mu}$ hours in all regions. We shall add the further

restrictive assumption that no inter-regional ambulance dispatch occurs. While this does not accurately reflect the behavior of either existing or possible future systems, by making this approximation we are able to make rough estimates of the number of ambulances to allocate to each region.

Under these assumptions it follows that the Expected Response Time in region j, $E[R_j]$ is equal to the sum of the Expected Dispatch Delay in region j, $E[D_j]$ plus the Expected Travel Delay in region j, $E[t_{NR_j}]$.

$$\text{i.e., } E[R_j] = E[D_j] + E[t_{NR_j}] \qquad (4.10)$$

for j=1,2,...,n if there are n regions.

The city-wide expected Response Time, E[R] is simply the weighted sum of the expected Response Time in each of the, say, n regions.

$$E[R] = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda} E[R_j] \qquad (4.11)$$

where: λ_j is the mean rate at which emergencies are generated in region j

λ is the mean rate at which emergencies are generated in the whole city

$$\lambda = \sum_{j=1}^{n} \lambda_j$$

The allocation of ambulances to the n regions now proceeds very simply. We allocate ambulances one-by-one, and in each case the allocation is to that region that produces the greatest reduction in the city-wide Expected Response Time given by 4.11. This allocation procedure continues until either there are no ambulances left to allocate, or we

have satisfied some criterion relating to the expected Response Time (either city-wide or regional). We shall demonstrate the method by means of a very simple example in Section 4.5. We pause at this point to summarize the assumptions made in this chapter, which are listed in Table 4.2.

4.5 A Simple, Numerical Example

For the purposes of this example we shall use the hypothetical city shown in Figure 4.7, for which all the assumptions of Table 4.2 hold. We assume that after consideration of the demand pattern in the city, there exist 3 distinct regions, across each of which the mean rate of call generation during any time period is approximately constant. Assuming that there are N ambulances available in the city, our task is to allocate them to the 3 regions so that the city-wide Expected Response Time is minimized.

For any particular distribution of the N ambulances among the three regions, say n_1 in region 1, n_2 in region 2 and n_3 in region 3 we can calculate the expected Response Times $E[R_1]$, $E[R_2]$, $E[R_3]$. The city-wide Expected Response Time is then simply

$$E[R] = \sum_{j=1}^{3} \frac{\lambda_j}{\lambda} E[R_j] \qquad (4.12)$$

Since the mean service rate μ is the same everywhere in the city, 4.12 can be rewritten

Table 4.2: Assumptions in Ambulance Allocation Model

1. The city or town to which the allocation is to be made is divided into n regions, in each of which calls arrive in a Poisson manner at a constant average rate. The average rate in region j is denoted by λ_j calls per hour.
2. The total time to service an emergency call is independently drawn from the negative exponential distribution with mean $1/\mu$ hours for any call in the city.
3. Ambulances are assigned to each of the n regions, randomly distributed throughout each region, and respond only to calls in the region of assignment. In every case, the nearest ambulance in the region is dispatched to the incident.
4. A call is delayed when all of the ambulances in the region are unavailable for dispatch. Delayed calls form a queue and are answered as ambulances become available on a first-come, first-served basis.
5. There are no external sources of ambulances. All calls are serviced by one of the ambulances allocated in the city.
6. The streets in the city are mutually perpendicular, and the average travel speed is the same across each region.

Figure 4.7: Hypothetical City for Numerical Example

Region 3 λ_3 = 2 calls/hour Area: 4 sq. miles	Region 1 λ_1 = 4 calls/hour Area: 2 sq. miles
	Region 2 λ_2 = 3 calls/hour Area: 2 sq. miles

Expected Service Time: 30 minutes in each region

Average Speeds: 10 mph in Regions 1 and 2
15 mph in Region 3

The city above has characteristics similar to a densely-populated city of about 1 million people.

$$E[R] = \sum_{j=1}^{3} \frac{\rho_j}{\rho} E[R_j] \qquad (4.13)$$

where $\rho_j = \frac{\lambda_j}{\mu}$

$$\rho = \sum_{j=1}^{3} \rho_j$$

The desired allocation is that which among all the possible n_1, n_2 and n_3 minimizes 4.13. Stated mathematically we are seeking

$$E[R^*] = \min_{n_1,n_2,n_3} \frac{1}{\rho} (\rho_1 E[R_1] + \rho_2 E[R_2] + \rho_3 E[R_3]) \qquad (4.14)$$

To simplify our calculations we shall work with the expression for $\rho E[R^*]$, which clearly does not affect the optimal allocation. We shall invoke the Optimality Principle of Dynamic Programming[25] and allocate the ambulances sequentially, assigning an ambulance at stage k to the region among the three which causes the largest reduction in $\rho E[R^*]$ found at stage k-1. We carry out the allocation in N stages; and the resultant distribution has the minimum city-wide Expected Response Time of all distributions with N ambulances.

Using the appropriate values for the city given in Figure 4.7 we can calculate, for each region, the Expected Delay Time, the Expected Travel Delay and the Expected Response Time. These are tabulated in columns 2, 3 and 4 of Table 4.3. In column 5 we enter the Expected Response Time for the region weighted by the demand intensity in the region. Deducting the weighted Expected Response Time for region j with n_j ambulances ($n_j \geq 1$) from the weighted Expected Response Time with (n_j-1) ambulances

Table 4.3: Calculations for Ambulance Allocation Example; Region 1

Region 1: $\frac{1}{\mu_1}$ = 30 mins.; λ_1 = 4 calls/hour; ρ_1 = 2; Av. Travel Speed = 10 mph

No. of Ambulances n_1	Expected Delay Time $E[D_1]$ mins.	Expected Travel Time $E[t_{NR1}]$ mins.	Expected Response Time $E[R_1]$ mins.	Weighted Exp. Response Time $\rho_1 E[R_1]$ mins.	Change in Wtd. Exp. Resp. Time $\Delta(\rho_1 E[R_1])$ mins.	Order of Ambulance Allocation
1	∞	-	∞	∞	-	1
2	∞	-	∞	∞	-	2
3	9.4	4.8	14.2	28.4	∞	6
4	2.6	3.9	6.5	13.0	15.4	9
5	0.6	3.4	4.0	8.0	5.0	12
6	0.0	2.9	2.9	5.8	2.2	

Table 4.3 (Cont.): Region 2; Region 3

Region 2: $\frac{1}{\mu_2}$ = 30 mins.; λ_2 = 3 calls/hour; ρ_2 = 1.5; Av. Travel Speed =10 mph

No. of Ambulances n_2	Expected Delay Time $E[D_2]$ mins.	Expected Travel Time $E[t_{NR2}]$ mins.	Expected Response Time $E[R_2]$ mins.	Weighted Exp. Response Time $\rho_2 E[R_2]$ mins.	Change in Wtd. Exp. Resp. Time $\Delta(\rho_2 E[R_2])$ mins.	Order of Ambulance Allocation
1	∞	-	∞	∞	-	3
2	>15.0	5.4	>20.0	>35.0	∞	5
3	4.0	4.2	8.2	12.3	22.7	8
4	0.9	3.8	4.7	7.1	5.2	11
5	0.2	3.0	3.2	4.8	2.3	

Region 3: $\frac{1}{\mu_3}$ = 30 mins.; λ_3 = 2 calls/hour; ρ_2 = 1; Av. Travel Speed = 15 mph

n_3	$E[D_3]$ mins.	$E[t_{NR3}]$ mins.	$E[R_3]$ mins.	$\rho_3 E[R_3]$ mins.	$\Delta(\rho_3 E[R_3])$ mins.	Order of Allocation
1	∞	-	∞	∞	-	4
2	12.0	4.4	16.4	16.4	∞	7
3	1.4	3.5	4.9	4.9	11.5	10
4	0.2	3.1	3.3	3.3	1.6	

gives us the reduction in $\rho E[R^*]$ at stage (n_j-1) resulting from the allocation of one vehicle to j when there are already (n_j-1) vehicles assigned to j. Finally we list in column 7 the order in which the ambulances are allocated, starting from 1.

For the sake of the example we shall assume that we have up to 12 ambulances available for allocation in the city. Inspection of Column 5 in Table 4.3 shows that we must allocate at least 3 ambulances to Region 1, and 2 ambulances to both Regions 2 and 3 in order to avoid theoretically infinite response times. Therefore, the first 7 ambulances are allocated in this manner. The allocation of the eighth ambulance to Region 2 results in the largest reduction in $\rho E[R^*]$ at stage 7. (22.7 minutes in Region 2, as opposed to 15.4 minutes in Region 1 and 11.5 minutes in Region 3.) The ninth ambulance is assigned to Region 1 and causes a reduction in $\rho E[R^*]$ at stage 8 of 15.4 minutes. The tenth vehicle is assigned to Region 3, and so on. The final allocation of ambulances is:

Region 1: 5 ambulances

Region 2: 4 ambulances

Region 3: 3 ambulances

By substitution from Table 4.3 we see that

$$E[R^*] = \frac{1}{2.0 + 1.5 + 1.0} [8.0 + 7.1 + 4.9]$$

$$= 4.5 \text{ minutes}$$

If budget constraints limited the number of available ambulances to 10, then the two ambulances allocated last [Nos. 11 and 12 in column 7 of Table 4.3] would be discarded. The resulting allocation is:

Region 1: 4 ambulances

Region 2: 3 ambulances

Region 3: 3 ambulances

As we would expect the city-wide expected Response Time rises, and we now find

$$E[R^*] = 6.7 \text{ minutes}$$

In the event that budget constraints were relaxed and another ambulance purchased, column 7 in Table 4.3 indicates that it should be assigned to Region 2. The new city-wide Expected Response Time is

$$E[R^*] = 3.9 \text{ minutes}$$

Instead of seeking to minimize city-wide Expected Response Time, we could instead have chosen another criterion. For example we might seek that allocation which requires the minimum number of ambulances that results in an Expected Response Time in each region of less than 5 minutes. Inspection of column 4 in Table 4.3 indicates that the required allocation is

Region 1: 5 ambulances

Region 2: 4 ambulances

Region 3: 3 ambulances

In summary, this example has demonstrated how an ambulance administrator might obtain rough guidelines to allocating ambulances around his city, taking into account the cost constraints and the consequences to the patient in terms of the delay experienced. Since the modelling of spatially distributed service systems is only in its infancy, the measures used here are gross approximations. Nevertheless by making use

of these simple models based on easily obtained parameters, a rational "first-cut" is available for distributing vehicles. This is definitely an improvement on procedures in all but a few cities where rules of thumb, tradition and convenience govern the allocation of these resources.

For the reader's convenience we include as Table 4.4 a list of the symbols used in this chapter.

4.6 Chapter Summary

We have succeeded in developing and using two simple analytical models that can help an administrator in allocating emergency ambulances to a city or town. These have both been predicated on the not unreasonable premise that the sooner an ambulance arrives on the scene of the emergency, the better will be the patient's chances of recovery.

The models have two important outputs that can be used by the emergency ambulance administrator:

(i) Insights into System Operation: The very process of developing analytical models has provided us with a number of insights into aspects of the emergency ambulance system. In this chapter we have seen that low utilization of vehicles, especially for small services, is an inevitable concomitant of a high level of service when calls arrive randomly. The high cost of providing an improved service level at the margin prompts a search for cheaper ways of servicing the calls causing saturation. Section 4.3.2 demonstrated the advantages of dispersing ambulances as opposed to having them at a single location.

Table 4.4: Summary of Symbols Used in Chapter 4

A:	Area
$E[D_c]$:	Conditional Expected Dispatch Delay
E[D]:	Unconditional Expected Dispatch Delay
$E[t_{NC}]$:	Expected Travel Delay with N centrally located ambulances
$E[t_{NK}]$:	Expected Travel Delay with N randomly located ambulances
$E[R_j]$:	Expected Response Time in Region j
E[R]:	City-wide Expected Response Time
$E[R^*]$:	Minimum city-wide Expected Response Time for a fixed number of ambulances allocated to n regions
m,n:	Dimensions of a rectangular city
n_j:	Number of ambulances allocated to Region j
P_i:	Steady-state probability that i patients will be undergoing and awaiting service
Q_N:	Steady-state probability that a system with N ambulances will be saturated
U:	Ambulance Utilization
v:	Travel Speed
λ:	Mean rate of call arrival
μ:	Mean rate at which calls are serviced
ρ:	Ratio of λ to μ

Finally the worked example in Section 4.5demonstrated the importance of the Travel Delay in determining the allocation of ambulances [see Table 4.3]. It is totally inadequate to consider Dispatch Delay alone and to assume that the time for the ambulance to reach the patient is negligible.

(ii) A Methodology for Ambulance Allocation: Very gross estimates on the one hand, extremely cheap and easy to use on the other, these analytical models guide the ambulance administrator in locating and allocating vehicles to various parts of the city. They guide allocations by forcing the recognition by the administrator of the constraints under which he is operating, i.e., financial limitations and intolerable delays to patients. Whereas computer simulations may produce more accurate allocations, these are not widely available and are unlikely to be in very general use in the near future. Under these circumstances the simple analytical model is the only resort of the administrator whose aim is an ambulance service tailored to take account of the number and location of calls. By adopting the procedure outlined in Table 4.5 the ambulance administrator has a tool for:

(a) Making allocations of ambulances around the city that reflect demand for emergency service and the topology of the region.

(b) Investigating the consequences of alternative strategies, especially with regard to allocating ambulances during different working shifts.

(c) Making budget predictions, and planning to meet anticipated changes in demand.

Table 4.5: Procedure for Using the Analytical Models in Chapter 4

1. Measure or estimate the mean call arrival rates across the city at different times of the day and the week.

2. Divide the city into regions where, for socio-economic and other reasons, the mean arrival rates are approximately invariant spatially.

3. Measure or estimate the mean time to service an emergency call.

4. Calculate the Expected Dispatch Delay and the Expected Travel Time in each region for different allocations of ambulances to the regions. Hence calculate the Expected Response Time.

5. Combine the Expected Response Times into a city-wide Expected Response Time for various allocations of ambulances.

6. Select that allocation that best satisfies budget constraints with an Expected Response Time that is acceptably low.

CHAPTER 5: PRIMARY AND SECONDARY AMBULANCES

5.1 Introduction

In Chapter 4 we proposed and modelled an emergency ambulance service in which a fixed number of ambulances were provided during a work shift to respond to the demand during that shift. A measure of the effectiveness of the system was the Service Level, defined to be the probability that the random patient encounters a dispatch delay. While the system described is essentially similar to many real-world systems, it is deficient in a number of respects:

(i) The utilization of the ambulances is extremely low for reasonably high levels of service (low probabilities of delay). See Figure 4.4.

(ii) It is extremely difficult to decide what constitutes an adequate level of service.

(iii) The system is inflexible. An unexpectedly low demand rate during a work shift results in idle crews, but no reduction in cost. An unexpectedly high demand rate on the other hand causes a drop in the level of service.

(iv) The system does not make use of all the ambulance facilities available in most cities and large towns.

In this chapter we shall consider the introduction of a secondary source of ambulances. As before we shall have a primary purveyor of

ambulance service, perhaps provided by the municipal government*, but in addition we shall provide support vehicles from another source. The presence of these secondary vehicles removes the need to provide sufficient primary ambulances to immediately respond to 95% or 99% of all calls. (Reference No. 5.) A result of doing this seems to produce a significant cost saving, even when a very generous policy of reimbursing the secondary operators is followed.

In the remaining sections of this chapter we develop a model that describes the situation when a secondary source of ambulances is introduced. From the model we can select the least-cost number of primary ambulances to meet a specific demand for service. By approximating the cost of operating an ambulance service by a linear function of the number of primary vehicles, we can compare the costs of systems with and without secondary ambulances.

To start with, we shall assume that there exists an inexhaustible supply of secondary ambulances. During the analysis it will become apparent that with high probability the inexhaustibility can be achieved with very few vehicles. After the analysis we shall discuss alternative sources of secondary ambulances.

5.2 The Primary-Secondary Models

We now postulate an emergency ambulance system in which there are

*We shall assume that either the service is provided free of charge or that patients are billed at a fixed rate that is independent of the mode of transport (primary or secondary).

two sources of ambulances. The primary source dispatches ambulances to all emergency calls as long as it has an ambulance available. The costs incurred by the primary purveyor are those of manning a fixed number of ambulances during each shift in the day, irrespective of the number of calls answered. We have indicated (in Chapter 2) that most of this cost is due to attendant salaries, and we shall therefore assume it to be a linear function of the number of ambulances in each shift. The secondary source functions as a reserve to the primary source, responding to calls arriving while all the primary vehicles are busy. We shall assume that the secondary purveyor is guaranteed payment on a pro rata basis by the operators of the primary ambulance service, i.e. secondary calls are charged to the primary operator at a fixed rate per call.

We can model the system described above in very much the same way as we modelled the saturation delay in section 4.2. Let us assume that in the time period under consideration calls for the emergency ambulance service arrive randomly in time at a constant average rate, λ calls per hour. We further assume that the time to service any call is drawn from the general distribution with finite mean $\frac{1}{\mu}$ hours. We understand that the service time includes time spent travelling to the scene of the emergency and to the hospital. Finally we assume that there are two sources of ambulances: a primary source with N manner vehicles and a secondary source with an unspecified number of vehicles. The primary ambulances respond to requests for emergency aid whenever a primary vehicle is available. The secondary ambulances respond only when requested to do so because no primary vehicle is free. We assume that

Table 5.1

Assumptions in Secondary Ambulance Source Model

(1) There exist two sources of ambulances to respond to calls for emergency service: a primary source which dispatches one of N primary ambulances to any call for which such a vehicle is available; and a secondary source which dispatches secondary ambulances to calls that arrive when all the primary ambulances are busy.

(2) Calls arrive in a Poisson manner at mean rate λ per hour. Service times are drawn from the general distribution with mean $1/\mu$ hours, for both types of ambulance.*

(3) The probability that no secondary vehicles are available is negligibly small.

(4) The costs of operating the primary arm of the service can be approximated as a linear function of the number of primary vehicles.

(5) The secondary source is paid on a pro rata basis.

*It is unlikely that the mean service time will be the same for both primary and secondary vehicles. The assumption is made only to facilitate exposition and can be dropped when necessary.

the probability that no secondary vehicles are available is negligibly small. All these assumptions are summarized in Table 5.1.

Because the calls arrive in a Poisson manner, the probability of k calls arriving in time T is

$$P(k;T) = \frac{(\lambda T)^k e^{-\lambda T}}{k!} \qquad T > 0; \quad k=0,1,2,... \qquad (5.1)$$

The expected number of calls in time T is given by

$$E[k;T] = \lambda T \qquad (5.2)$$

After the system has been in operation for a while the primary source will be in one of N+1 "states" to which we may attach a steady-state probability. For state i, the steady-state probability P_i is the probability that at any random time i of the N primary ambulances are "servicing" emergency calls. (i=0,1,...,N.) We recall that there can be no states greater than N because calls arriving when the primary source is in state N are answered by secondary ambulances. Because there are N+1 mutually exclusive, collectively exhaustive states we can write

$$\sum_{i=0}^{N} P_i = 1 \qquad (5.3)$$

In the situation where we have calls arriving at the primary in a Poisson manner at mean rate λ calls per hour and negative exponential service times with mean $\frac{1}{\mu}$ hours, it is easily shown that

$$P_i = \frac{\frac{(\lambda/\mu)^i}{i!}}{\sum_{i=0}^{N} \frac{(\lambda/\mu)^i}{i!}} \qquad (5.4)$$

It is a much more difficult task to show that the same result holds when service times are drawn independently from a general distribution with finite mean $\frac{1}{\mu}$. We shall not carry out the proof, but the interested reader is referred to Reference No. 19.

Using 5.4, it follows that the probability that all N primary ambulances are unavailable is simply

$$P_N = \frac{\frac{(\lambda/\mu)^N}{N!}}{\sum_{i=0}^{N} \frac{(\lambda/\mu)^i}{i!}} \qquad (5.5)$$

Over a long period of time T, the primary source spends a time $P_N T$ in state N. All calls arriving during this time are answered by a secondary ambulance. Therefore, the expected number of calls answered by the secondary source in time T is

$$E[k_s;T] = \lambda P_N T$$

In unit time (one hour) the secondary ambulances answer an average λP_N calls. Correspondingly, in the same time, the primary source answers $\lambda(1-P_N)$ calls on average.

5.3 The Expected Cost of Operating the System

With the aid of these simple results we can estimate the expected cost of operating the primary-secondary system. We shall assume that the cost of providing the primary service is directly proportional to the number of primary ambulances allocated to a shift. Let the primary

ambulance cost be c_p per vehicle per hour. This includes all the costs associated with the vehicle and its crew. In Chapter 2 we indicated that these costs account for about 80% of the total cost of providing an ambulance service, and so they do represent a reasonable approximation to total cost.

We assume that the secondary source gets paid at the fixed rate of c_s per call answered. Since the average time to service a call is $\frac{1}{\mu}$ hours, the average fee to the secondary source per hour spent working is

$$c_s' = \mu c_s \qquad (5.7)$$

Since we know μ, it is easy to transfer from one cost to the other.

From all the assumptions we have made we can now calculate the expected cost per hour of operating the system we have modelled:

(i) The primary cost, irrespective of the number of calls answered is simply $N.c_p$ per unit time

(ii) The expected secondary cost is the product of the expected number of calls answered in unit time and the cost per call, and equals $\lambda.P_N.c_s$ (from 5.6)

(iii) The total expected cost per unit time is therefore:

$$C = N.c_p + \lambda.P_N.c_s$$

$$\text{or} \qquad C = N.c_p + \frac{\lambda}{\mu} P_N.c_s' \qquad (5.8) \quad \text{(from 5.7)}$$

Since c_p and c_s' are in the same units we can write

$$c_s' = r.c_p$$

where r is the average ratio of the secondary cost per hour of operation to the

primary cost per hour. We can therefore finally write equation 5.8 as

$$C = c_p[N + r\rho\, P_N] \qquad (5.9)$$

$$\text{where} \quad \rho = \frac{\lambda}{\mu}$$

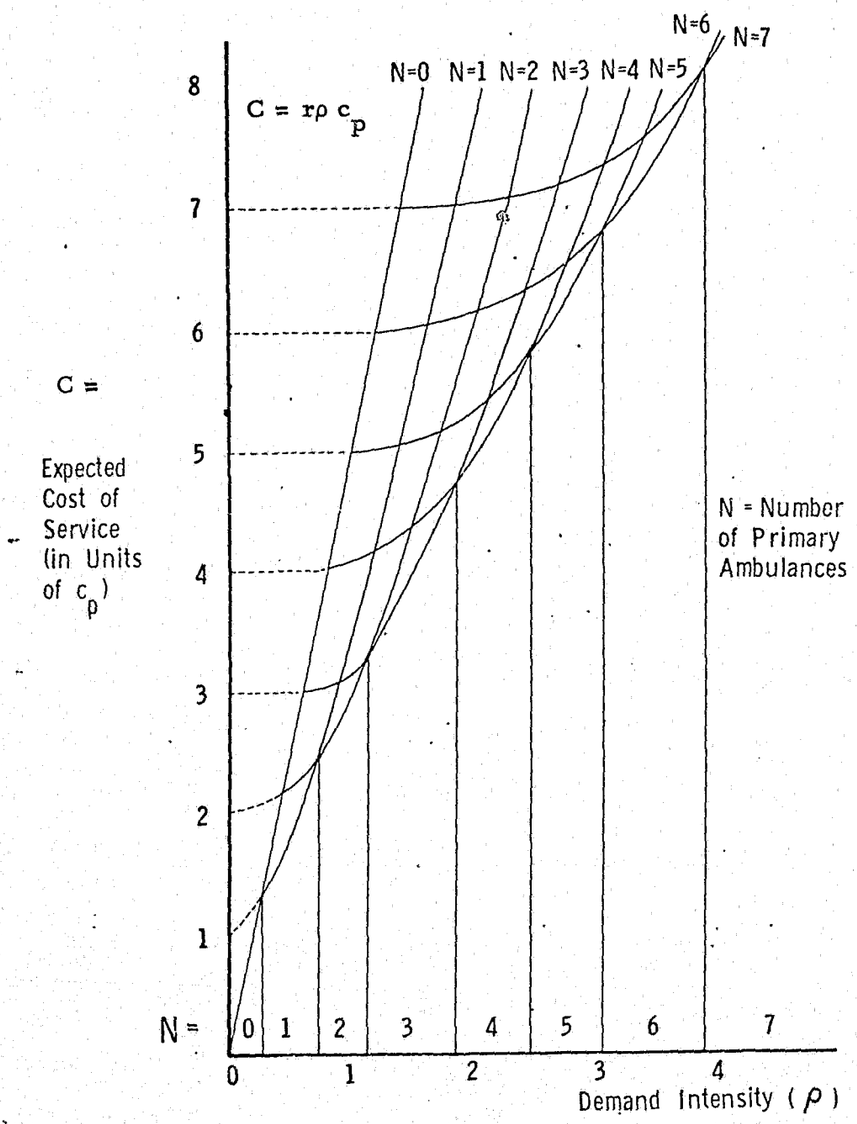
Equation 5.9 has been graphed in Figures 5.1 and 5.2 as a function of ρ with the number of primary ambulances, N as a parameter. We have arbitrarily set r=5. The curves representing equation 5.9 combine to produce a minimum cost envelope, so that at each value of ρ it is possible to specify the number of primary ambulances that result in the minimum expected cost.

5.3.1 The Value of r

The reader should note that unless r is greater than unity it is always more economical to provide the whole service using only "secondary" ambulances. Even if r is greater than unity, there will be a critical value of ρ below which it would be more economical to operate with the secondary ambulances alone rather than to introduce a primary service. We offer the following simple proof of these statements:

From equation 5.9, the expected cost of operation without a primary source is simply $C = c_p r\rho$. ($P_N = P_o = 1$, because there are no primary vehicles.) The expected cost of operation with one primary ambulance is $C' = c_p + r\rho\, P_1 c_p$ (from equation 5.9) In this case the primary ambulance can be in one of two states: available (with probability P_o) or answering a

FIG. 5.1: COST OF PROVIDING AN E.A.S. WITH PRIMARY AND SECONDARY SERVICES (r = 5)



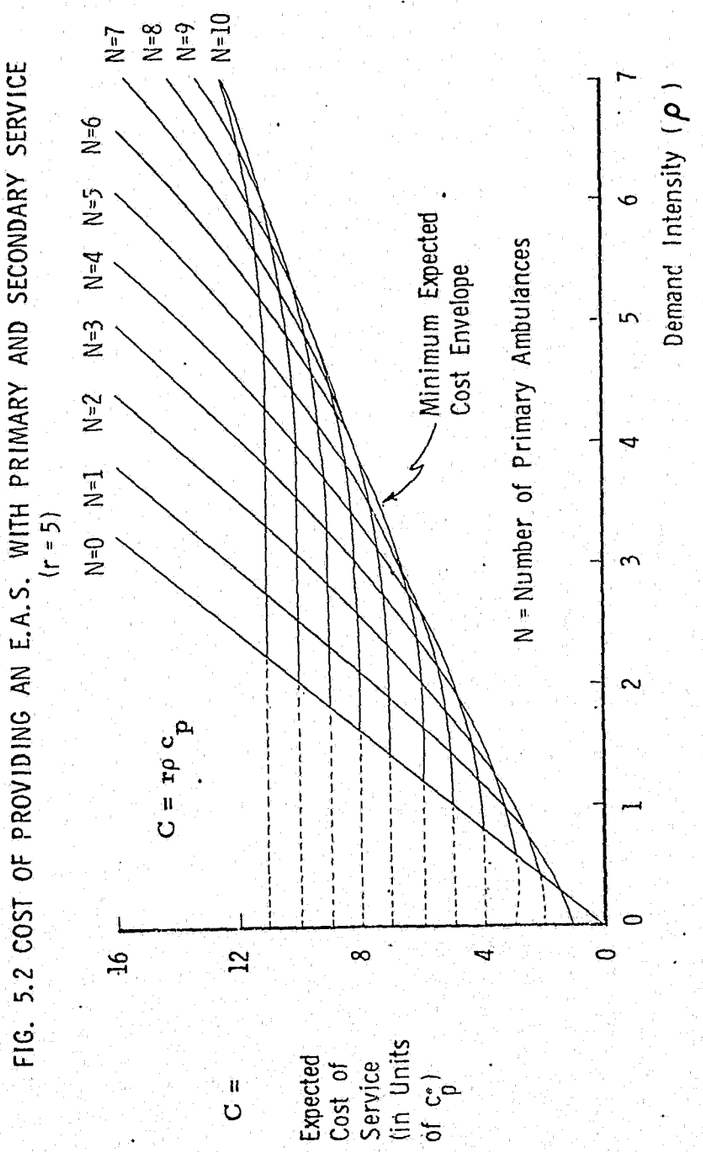


FIG. 5.2 COST OF PROVIDING AN E.A.S. WITH PRIMARY AND SECONDARY SERVICE (r = 5)

call (with probability P_1). From equation 5.4 it follows that $P_1 = \frac{\rho}{1+\rho}$.

Clearly it becomes economical to employ a primary ambulance when the expected cost with that ambulance is less than the expected cost without it, i.e. when $C' < C$

$$\text{when} \quad c_p r\rho > c_p\left[1 + \frac{r\rho^2}{1+\rho}\right]$$

From the above we find that a primary ambulance should be introduced only if $r > 1 + \frac{1}{\rho}$. Therefore, if $r \leq 1$ it never pays to use primary vehicles. The whole service can be more economically provided by ambulances paid on a pro rata basis. Even if $r > 1$, if $\rho \leq \frac{1}{r-1}$ it is economical to use only secondary vehicles to provide the service. Reference to Figures 5.1 and 5.2 shows that for small values of ρ no primary ambulances are involved.

To summarize, if a secondary source can be found that will provide ambulance service on a pro rata basis such that $r \leq 1$, then there is no reason to consider providing additional primary ambulances. Where the secondary to primary ratio is greater than unity (a most probable event), if the demand intensity ρ is small then it will still be economical to operate with secondary vehicles alone. This is an obvious solution to the problem of providing emergency service in small towns where the demand is low.

5.3.2 The Average Cost per Call Ratio, k

The ratio r relating primary cost per hour to secondary average cost per hour worked is not a parameter that is easily conceptualized. The average ratio of the primary cost per call to the secondary cost per call, k might be more attractive. Unfortunately k is a function of ρ and N, and is therefore not a constant. Nevertheless it is possible to derive a simple general relationship defining k, and from it to draw an interesting result. From previous work we know:

The average cost per call for the secondary ambulances is c_s

$$c_s = \frac{c'_s}{\mu} = \frac{r\, c_p}{\mu}$$

The average cost per call for the primary ambulances is given by

$$c'_p = \frac{N\, c_p}{\lambda(1-P_N)}$$

$$k = \frac{c_s}{c'_p}$$

$$= \frac{r\rho(1-P_N)}{N}$$

$$r\rho - r\rho P_N - N = (k-1)N$$

$$[r\rho - \frac{C}{c_p}] = (k-1)\, N \qquad \text{from equation 5.9}$$

$$k = 1 + \frac{1}{Nc_p}\, r\rho c_p - C \qquad (5.10)$$

The quantity $(r\rho c_p - C)$ can be obtained from Figure 5.1 and is always positive when both primary and secondary sources are

used. From 5.10 it follows that under these circumstances $k > 1$.

Calculating values of k from Figure 5.1 for various values of ρ we find that for r=5, k varies in the range 1·5 - 2·5. Effectively we have constructed a service in which the secondary source is being paid a fee per call that is approximately twice the cost per call for the primary source. This is a deliberate effort to encourage participation in the emergency ambulance service by secondary sources by making it a financially attractive prospect.

5.4 Discussion of the Model

A major result of the modelling effort is the graph shown in Figure 5.1 (and 5.2) which allows the administrator of a municipal emergency ambulance service to allocate specific numbers of ambulances to work shifts given some knowledge of the expected demand. A particular example of the minimum cost envelope that results from the primary-secondary system when r=5 is presented in Figures 5.1 and 5.2. Along the abscissa in Figure 5.1 we indicate the range of ρ for which a particular number of primary vehicles affords the lowest cost of system operation.

In theory, armed with graphs like those in Figures 5.1 and 5.2 an administrator could allocate the number of primary ambulances needed in each shift to meet the demand (represented by ρ) at the lowest cost. Unfortunately, just as in the single ambulance source situation of Chapter 4, the average rate at which emergency calls arrive, λ, is unlikely to be constant over the duration of the shift. Therefore any

allocation of vehicles will only be "optimal" for a limited period of time. In this case, as we shall see below, the consequences are only in terms of costs and not in terms of the quality of service offered to emergency victims.

5.4.1 Advantages of the Dual-Source System

There appear to be a number of advantages to the system proposed in this chapter, which to the author's knowledge have not been fully exploited in any communities to date.

(i) In general it appears that the dual source system offers much greater flexibility than the single source system discussed previously. Reasonably wide fluctuations in the demand rate may be accommodated without causing patients serious delays. Conversely the consequences of an erroneous prediction in the expected demand in a shift would be less costly or less serious than in a single-service system. For example, when ρ is less than the value for which the allocation of primary vehicles was originally made, then the costs of operation fall because less use is made of the secondary vehicles. This is in contrast to the single source situation where costs are independent of the number of calls answered. On the other hand, when ρ increases above the level anticipated, all calls continue to be answered (under the assumption that

secondary ambulances are never depleted*) in the dual-source case. The costs of the service rise above those for an optimal allocation of vehicles at the new demand rate, however the level of service (0% dispatch delay) is maintained. In the single source situation (Chapter 4), as demand increases the level of service declines with a fixed number of ambulances. Because maintaining a high probability of immediate response is an important criterion, under a system with a fixed single source of ambulances, this requires the allocation of sufficient vehicles throughout the shift to meet the highest demand rate. This inevitably leads to a low-utilization, high-cost service.

(ii) Even more important than flexibility may be the absolute reduction in costs that result from answering marginal calls with a secondary ambulance. In Figure 5.3 we compare the costs of providing 3 levels of single-source service with the

---

*Assume that there are an infinite number of secondary ambulances and N primary ambulances. Then the probability of there being k secondary vehicles busy at one time is easily shown to be

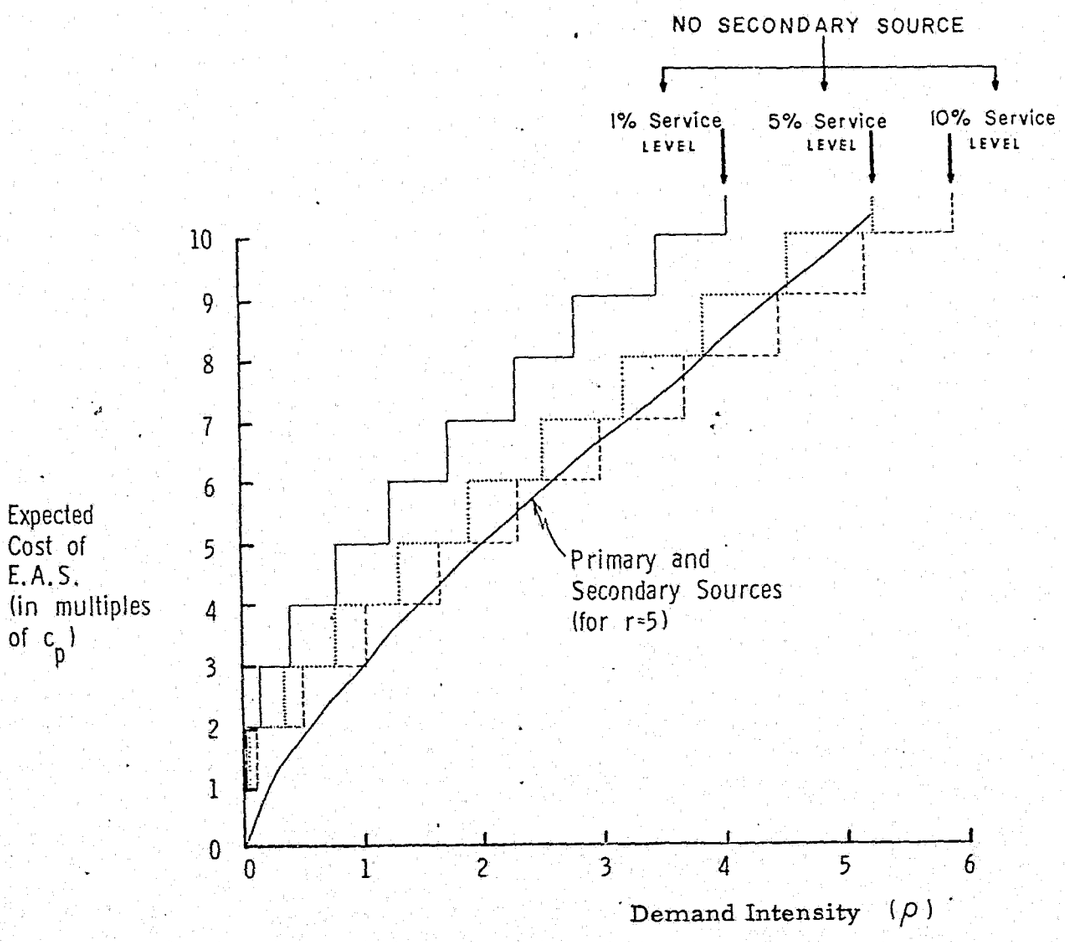
$$\Pr[k \text{ busy}] = \frac{(\rho P_N)^k e^{-\rho P_N}}{k!} \qquad k=0,1,2,\ldots$$

Therefore, the probability that more than one secondary vehicle is busy at a random time is

$$\Pr[k > 1] = 1 - (1 + \rho P_N)e^{-\rho P_N}.$$

For the situation when r=5 we can show by checking points on Figure 5.1 that this probability is about 6%. We therefore believe that our assumption about inexhaustible secondary vehicles is valid.

FIG. 5.3: COMPARISON OF COSTS FOR DUAL AND SINGLE SOURCE MODELS



dual-source service in which r=5. It appears that by paying a secondary source approximately twice as much per call as the primary source (see Section 5.3.2) we can provide a service with no dispatch delay at about the same cost as a single source operating at a 10% service level. The cost reduction is most significant for small demand intensity ρ. The reader should keep in mind that we have approximated the cost of providing the service as a linear function of the number of primary vehicles involved.

(iii) A final consideration is that the existence of a secondary source of ambulances allows the dispatcher some discretion in the vehicle he chooses to dispatch in response to a call. For example, at a point at which very few primary vehicles are free, the dispatcher has the option of sending a secondary ambulance to a call that he perceives as non-urgent, and of keeping primary vehicles in reserve for urgent calls. This strategy will in general prove to be more expensive than the simple system of our model, but it could provide better service to seriously injured people if the secondary service is of a lower standard than the primary. Alternatively, if there is an available secondary ambulance in the neighborhood of an incident, while the nearest primary vehicle is a considerable distance away, then dispatching the secondary rather than the primary vehicle would reduce the time the patient waits for the ambulance's arrival.

5.4.2 The Existence of Secondary Sources

So far we have assumed that sources of secondary vehicles exist without specifically examining what these might be. Potentially the most important source would be private ambulance companies engaged in the routine transport of patients between home and hospital. These groups have been at best reluctant participants in the emergency service because of the difficulty of providing high quality care at reasonable (to the patient) rates, and the problem of collecting payment after an emergency call. If the private companies could be guaranteed payment by the municipality at reasonably high rates for each emergency call undertaken, then it should be possible to persuade them to join the service and to maintain acceptable standards.

Another set of secondary ambulances to a primary municipal service could be the primary ambulances from neighboring communities. This situation is more complex because of the interactions among the communities, and would require regionwide coordination of emergency services among different communities that does not exist at present.

A third source of secondary ambulances are police station-wagons (currently used as primary vehicles in many communities) which could be used for less serious emergencies. On the whole they are unsuitable for emergency work, being very lightly equipped and poorly designed for en route emergency treatment. Better designed and equipped are the fire rescue wagons attached to the fire departments in most cities. Unfortunately the consequences of

the rescue wagon being unavailable when needed at a fire may be too serious to warrant its use as a secondary ambulance.

5.5 Conclusion

In this chapter we have developed and discussed a model of a particular type of municipal emergency ambulance operation. The model is primarily an aid to decision-making and would be used by administrators for

(i) Allocating primary ambulances to the daily working shifts, and updating these allocations as demand changes.

(ii) Calculating budget requirements for the future operation of the service, and making decisions about the purchase of new equipment.

(iii) Changing the primary operation in response to changes in the fees required by secondary operators.

The primary-secondary system explored in this chapter appears to have the following advantages (revealed in part by the model) over the single-source service:

(i) A reduced cost at a higher level of service.

(ii) Much greater flexibility, resulting in lower costs when demand drops within a working shift, and the maintenance of the level of service when demand rises.

(iii) The specially large cost-reduction over single-source operations when demand is low, i.e. for small services.

(iv) The voluntary incorporation of the private operator into the community emergency ambulance service, which would encourage higher operating standards, and would improve coordination during disaster situations.

One premise of this modelling effort was that an increased use of already available facilities could produce an improved service. This document has only considered better use of facilities inside one community. The next step is to improve the utilization of the facilities that exist within a region consisting of a number of communities. This too could be aided by a modelling effort that would indicate the allocation of ambulances to specific communities and establish decision rules for inter-community ambulance dispatch and the relocation of ambulances among communities in response to the dynamics of the demand for emergency service.

CHAPTER 6: SUMMARY AND CONCLUSION

6.1 Summary of the Document

In Chapter 1 we outlined the rise of the emergency ambulance service from obscurity to somewhat depressing prominence. We pointed out the great diversity of purveyors and indicated the complete lack of any central responsibility for the emergency ambulance function. In rural areas we found that morticians and private companies provided a service that in general does not satisfy standards set by the medical community with regard to attendant training and vehicle equipment. We also found that they operate under great financial strain in their efforts to collect adequate payment from emergency victims alone. In the larger metropolitan areas we saw that the emergency ambulance service is more often provided by a municipal agency free of charge to the emergency victim.

The establishment of legal standards of ambulance operation helped to define for the operator what some aspects of a good service might be like, without contributing significantly to their attainment. The first response to the promulgation of these standards, and to the rise in ambulance attendant salaries was the exodus from the emergency service of the morticians and some of the private companies. This in turn led to the recognition that the emergency ambulance "problem" was, at least in part, a financial one. Furthermore it became apparent that in order to

improve the quality of the service dedicated professional attendants needed to be attracted. It was quite clear that the minimum wage was not going to be sufficient inducement.

Through the literature a new awareness of the importance of the emergency ambulance service arose. There were positive responses from municipal governments to medical pressure for improved services. This then generated a new dilemma: how did one go about designing an improved service, and how much would it cost to operate? Standards had been set specifying attendant training, vehicle design and equipment, and first-aid procedure. Personnel, who were apparently qualified to make good ambulance attendants, notably army medical corpsmen, were available and could be attracted into the service at the right salary. The problem facing the administrator was how many of these relatively high-cost ambulance-attendant units should be allocated to which parts of the city to provide the most effective emergency service.

Our approach to the problem began with the division of the emergency ambulance service system into three component sub-systems: communications, medical services and transportation. Isolating the last resulted in the proposition of the ambulance response time as a measure of system performance. The division of the response time into two components, the dispatch delay and the travel delay, allowed us to develop mathematical models that related the number of ambulances assigned to the city, and their location, to the expected response delay. A number of assumptions that qualify the value of these models had to be made and these are discussed in Chapter 4. Nevertheless, it is a relatively simple matter

for an ambulance service administrator armed with data on the average rate of call generation, mean service times, and ambulance speeds to estimate the mean response time to an emergency call for a specific number of ambulances. The procedure in Chapter 4 gives order of magnitude estimates of the minimum expected response time for a specific number of ambulances allocated to various regions in the city when no inter-region dispatching is considered. Alternatively the procedure allows the estimate of the minimum budget required to attain a particular low expected response time.

Estimates of the utilization of the ambulances in services providing reasonably low response times are found to be very small - an inevitable characteristic of a service in which a high probability of immediate response is desired. Low utilization of ambulances implies a high cost to the user and a consequent fee that may appear to be disproportionately large for the service being provided.* This was one of the major causes of the financial difficulties under which private companies labored, and efforts to cut costs were at least partially responsible for the poor service provided. Fortunately a large number of communities, most of the big cities included, have realized that the benefits of an emergency ambulance service with the capability of immediate response are indirectly received by everybody in the community, not just by those who have the

---

*Assume that the mean service time is 30 minutes, that two attendants are required per ambulance, and that they are paid $5 per hour (the same salary as policemen and firemen). If attendant salaries are 80% of total cost, and the ambulance operates at a 25% utilization, then the fee per patient to cover costs is $5 \times 2 \times \frac{4}{2} \times \frac{100}{80} = \$25$. We recall from Chapter 2 that this is a good approximation to actual costs.

need to use it. Therefore the emergency ambulance service is increasingly being provided at community expense, or at least heavily subsidized by the community.

With emergency ambulance utilization rates in the range 25-35%, there would seem to be ample room for improvement. In Chapter 5 we attempted on the one hand to produce a more realistic model of the service, and also to examine a method of reducing costs by making better use of existing resources. We saw that by using existing private ambulance companies as secondary sources of emergency vehicles, the cost of providing the service could be reduced over that when no secondaries were available, and that in addition a higher level, more flexible service resulted. Effectively we eliminated the Dispatch Delay of Chapter 4, although the Travel Delay remains, and ambulance allocations would have to be made subject to acceptable Travel Delays.

6.2 Areas of Future Research

Research in the emergency ambulance field will not unreasonably be directed toward the improvement of the care delivered to the patient by each of the component sub-systems of the service, and to the reduction of costs. Let us begin by reviewing quickly a few strategies by which costs might be reduced.

6.2.1 Cost-Reduction Strategies

We have already investigated one cost-reduction strategy, that of introducing a secondary source of ambulances. While there is no

reason to believe that the secondary vehicles should necessarily come from private companies, they appear to offer the best alternative in terms of the quality of service they could provide. Other cost-reduction strategies that need to be researched are:

(i) Matching Response to Demand: Ideally we should like to assign just those ambulances needed to meet the demand for service. In practice assignments are made for the duration of a shift. Since the demand varies among shifts the staffing level will vary at different shifts. In order to accomplish an equitable workload, a schedule will need to be developed that ambulance attendants find acceptable. This apparently simple problem is yet to be resolved by police departments in this country.

(ii) Assigning Priorities to Emergency Calls: Because of the great range in the urgency of calls received by the emergency service, and because of the high cost of providing immediate response, it would be a great boon to system operation if low priority calls could endure a non-crucial delay when the service was close to saturation. We saw that the provision of secondary ambulances gave the dispatcher the option of sending one of these vehicles to a call that he suspected was of low priority. There needs to be developed a set of options for the

dispatcher in responding to calls of different priority (i.e., send primary ambulance, send secondary ambulance, send police station-wagon, delay response, etc.) and a set of guidelines that relate the nature of the response to the state of the system.

(iii) Combining Services: Figure 4.4 indicates that the utilization of ambulances (and hence the efficiency of system operation) increases as the demand intensity increases. This suggests that we might try to aggregate the demand rate in some fashion. One method might involve combining the emergency ambulance service with the fire rescue arm of the fire service. A second alternative would involve the combination of a number of local independent emergency ambulance services into one regional emergency service. This approach allows, in addition, the use of helicopters and special duty vehicles whose cost would prohibit their incorporation into one small service. A prerequisite for this type of service would be the establishment of a regionwide emergency communications network, and a very careful analysis of the points at which the emergency vehicles would be located. If the region were particularly large, there might be an advantage to relocating ambulances during the day in response to spatial changes in demand rate.

All of the above places a heavy burden of responsibility on the dispatcher whose task has been compounded by a larger region, a possible increase in his response options, and possibly a changing set of ambulance locations.

6.2.2 Other Areas of Research

In addition to some of the areas of research outlined above, there are at least three other important problems on which attention should be focussed which do not relate to costs:

(i) The collection of data and the execution of scientific experimentation to explain the generation of emergency patients. There are wide variations in the number of emergency ambulance patients per capita generated in different cities (see Chapter 2). This is only partly explained by accounting procedures. There are socio-economic and other factors involved, and by understanding these, and perhaps by providing more comprehensive preventive health care through public hospitals some people may be spared the opportunity of becoming emergencies. In addition, the more data that exists describing the generation of emergency calls the better can be the allocation of ambulances to respond to these.

The following is a list of data that might be collected for each emergency call answered by the

ambulance service:

Location of the call

By whom was the call reported

Time at which the call was received

Dispatch delay

Travel delay

Distance travelled by the ambulance

Time spent treating the patient at the scene

Time spent transporting the patient to hospital

Hospital to which the patient was taken

Ambulance dispatched: primary or secondary, local or other

Number of ambulances available when the call arrived

Nature of the emergency

Ultimate disposition of the patient

(ii) The improvement of medical services at the scene of the emergency, mainly through the use of more advanced communications technology. The installation of a communications system that allows a verbal exchange between hospital and ambulance and also the monitoring of the patient's condition at the scene by doctors some distance away (for example, taking an electro-cardiograph), will allow drastic treatment (for example, defibrillation) appropriate to an emergency situation but currently confined to a hospital environment.

(iii) The further development of analytical models, especially those related to vehicle location, as a cheap and quick aid to the assignment of vehicles in an urban environment. The data collection suggested under (i) could be used to validate, utilize and extend the models proposed in this document.

6.3 Conclusions

In this document we have examined three analytical models related to one aspect of the emergency ambulance service. The Dispatch Delay, the Travel Delay and the Dual Ambulance Source Models all relate to the Transportation component of the Emergency Ambulance System. The primary purpose of the document was to investigate the feasibility of using the modelling and analytical procedures of operations research in a "public system" like an ambulance service. Clearly the modelling procedures are useful, offering both fairly specific results (e.g. the superiority of the dual source system; the service time distribution) and more general insights (e.g. the low utilization of ambulances; the high cost of answering the marginal call). Because of the scarcity of data we have not validated these models satisfactorily, and therefore more work should be done in this direction.

From the point of view of the ambulance service administrator the most interesting output of this document is the procedure recommended for making allocations of ambulances to different parts of a city. This is the work of Chapters 4 and 5, and we summarize it briefly here.

(i) From data gathered in a similar basis as that described in Section 6.2, map out the rate at which calls are generated in different parts of the city at different times in the day. From this "map," divide the city into regions across each of which the call rate is approximately constant.

(ii) Record the mean call arrival rate, mean vehicle speed, and the mean service time in each region for every working shift in the week.

(iii) If there are no secondary ambulances available, calculate for each shift the Expected Response Time in each region for different numbers of ambulances. Then use the methods of Section 4.4 to allocate ambulances to each region such that the city-wide Expected Response Time is minimized, or some other criterion satisfied (e.g. regional expected response times below a specified minimum).

(iv) In the more likely event that secondary ambulances are available the situation becomes somewhat complicated. Allocate to each region the number of primary ambulances that minimizes operating cost in the region (Figure 5.2). Calculate the Expected Travel Time in each region. If these are satisfactory, then the solution is acceptable. If, on the other hand, there are some regions in which the Expected Travel Time is unacceptably large,

additional ambulances will need to be added to these regions. Departure from optimality necessitates a search for a Pareto-optimal solution in which the allocations to each region may change, so that acceptable Expected Travel Times are produced at minimum cost. Because of the number of alternatives the search might best be carried out by computer.

(v) Implement the suggested allocations and test for aberrations produced by particular conditions in the city not accounted for by the model.

As the cities and towns in the United States prepare to provide better equipped, rapidly responding emergency ambulances services, one of the first problems will be to prepare a detailed plan of system operation and anticipated costs. The modelling procedures described in this document can be used as a framework within which to describe alternative system configurations, to prepare budgets and thereby to remove some of the current obstacles to action.

APPENDIX A: AMBULANCE SERVICE TIMES

A.1 Introduction

In Figure 3.2, page 47 we defined a number of the parameters underlying the ambulance transport operation. One of these is the Service Time, the time between the ambulance being dispatched and its eventual return to duty after the patient has been delivered to the care of a hospital. Service Time is an important component of the models used in this document, but we have treated it in a cursory fashion, mainly because it does not lend itself to simple mathematical treatment.

In this appendix we shall explore one possible model of the Service Time. It will involve a number of simplifying assumptions, some of which will restrict the model's applicability to cities with particularly regular street plans. Comparisons with two ambulance operations in New York City suggest that despite the assumptions the model is not an unreasonable approach to describing the actual Service Time patterns of some real systems.

Our approach will be to consider the Service Time as the sum of three components. These are:

(i) The time to travel to the scene of the emergency.

(ii) The time to treat the patient at the scene.

(iii) The time to travel from the scene to the hospital.

By assuming that the ambulance is located at or near a hospital and that it transports the patient to that hospital, we may combine (i) and (iii)

above into Total Travel Time. The Service Time is now the sum of two independent times: Total Travel Time and Treatment Time. The Service Time distribution is then simply the convolution of the distributions of the two component times. We begin by modelling the distribution of the distance between an emergency and the ambulance when it is dispatched. We can then find the time taken to travel to and from the emergency (Total Travel Time). Finally we add the assumption that the time taken to treat the patient is distributed as the negative exponential.

A.2 Travel Distance

The context of our model is a city in which streets are laid out in the form of a grid. Emergency calls occur in the city and are distributed randomly in space. There are a number of hospitals located randomly throughout the city. Emergency ambulance service is provided by vehicles located at or near these hospitals, which respond to calls occurring in a region around the hospital (see Appendix B). We shall assume that trips out of the region correspond to a negligible fraction of the total number of ambulance dispatches. The ambulances transport patients back to the hospital associated with the region. For simplicity we can assume that the region is a rectangle with side m and n distance units long, parallel to the street grid. Locating the origin of a Cartesian coordinate system in the bottom left-hand corner of the region, we can represent any location by its x and y components. [See Figure A.1(a).]

Let the location of the hospital be at (x_H, y_H). We can denote the position of an incident by (x_i, y_i), where x_i and y_i are independent of

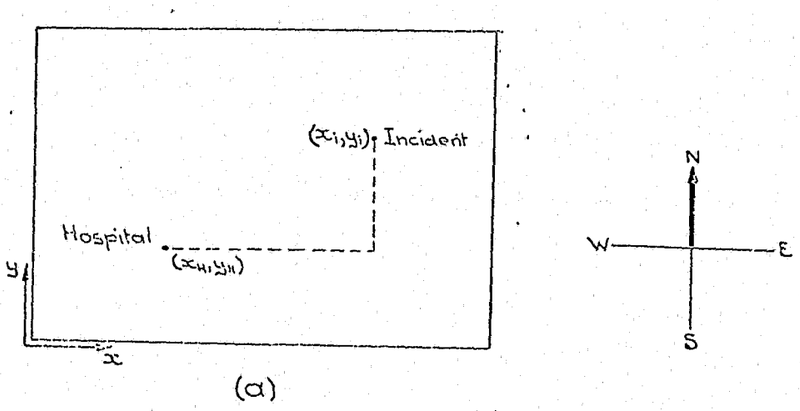


FIGURE A.1 THE LOCATION OF AN INCIDENT

each other, and are uniformly distributed over n and m respectively [Figure A.1(b) and (c)]. Approximating the travel distance between the hospital and the incident by the sum of the North-South and East-West distances, we can write:

$$\text{Travel Distance, } d = |x_H - x_i| + |y_H - y_i|$$

Setting $x_r = |x_H - x_i|$ and $y_r = |y_H - y_i|$.

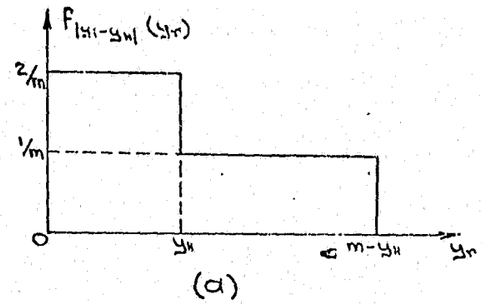
$$d = x_r + y_r \qquad (A.1) \qquad 0 \le x_r \le n \qquad 0 \le y_r \le m$$

A little thought will confirm that the density functions $f_{y_r}(y)$ and $f_{x_r}(x)$ are as shown in Figures A.2(a) and (b), and that the joint density function $f_{x_r y_r}(x,y)$ is described by Figure A.2(c). Specifically we can write that

$$f_{x_r y_r}(x,y) = \begin{cases} \frac{4}{mn} & 0 \le x_r \le x_H \quad 0 \le y_r \le y_H \\ \frac{2}{mn} & 0 \le x_r \le x_H \quad y_H \le y_r \le m-y_H \\ & 0 \le y_r \le y_n \quad x_H < x_r \le n-x_H \\ \frac{1}{mn} & x_H < x_r \le n-x_H \quad y_H < y_r \le m-y_H \end{cases} \qquad (A.2)$$

We can use Laplace Transform methods to show that the density function for the Travel Distance is:

$$f_d(d_o) = \frac{d_o}{mn}\,[4-2u_{-1}(d_o-x_H) - 2u_{-1}(d_o-n+x_H) - 2u_{-1}(d_o-y_H) - 2u_{-1}(d_o-m+y_H) + u_{-1}(d_o-x_H-y_H) - u_{-1}(d_o-n+x_H-y_H) + u_{-1}(d_o-m+y_H-x_H) + u_{-1}(d_o-m-n+x_H+y_H)] \qquad (A.3)$$



(b)

(c)

FIGURE A.2 DISTANCE BETWEEN AN INCIDENT AND THE HOSPITAL

The expected value of d, which is quite easily obtained from the Laplace Transform of $f_d(d_o)$, is found to be

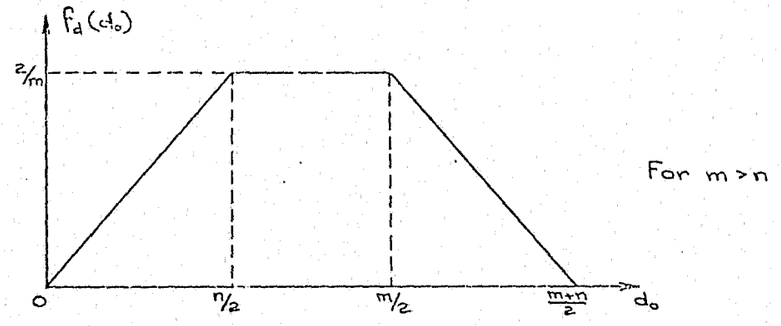
$$E[d] = \frac{m+n}{2} + \frac{y_H^2}{m} + \frac{x_H^2}{n} - x_H - y_H \qquad (A.4)$$

Let us now consider a few special cases of the above equation:

(i) A.4 is a minimum when the hospital is in the center of the region at $(\frac{n}{2}, \frac{m}{2})$. In this case $E[d] = \frac{1}{4}(m+n)$ (A.5)

The resulting density function may be derived from A.3 and is depicted below:

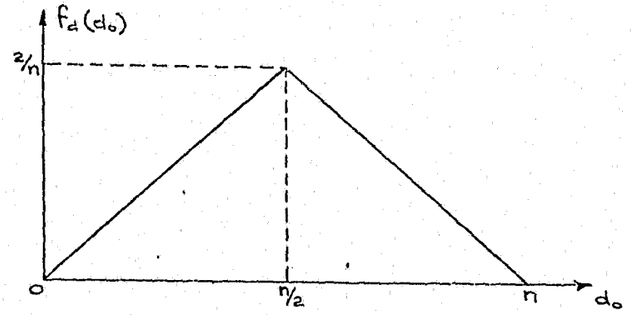
$$f_d(d_o) = \frac{4d_o}{mn}\left[1 - u_{-1}(d_o - \tfrac{n}{2}) - u_{-1}(d_o - \tfrac{m}{2}) + u_{-1}(d_o - \tfrac{m+n}{2})\right] \qquad (A.6)$$



(ii) If the region is a square (i.e. m=n), the A.4 becomes

$$E[d] = n + \frac{1}{n}(x_H^2 + y_H^2) - x_H - y_H$$

If we assume further that the hospital is centrally located, then $E[d] = \frac{n}{2}$, and the resulting density function is triangular in shape:



(iii) If the ambulance were not located at (x_H, y_H), but instead at some point uniformly distributed over the region, it would be necessary to take the expectation of A.4 over the possible positions (x,y). Therefore, with a randomly located ambulance A.4 becomes

$$E[d] = \frac{1}{3}(m+n) \qquad (A.7)^*$$

If the region is a square A.7 reduces to:

$$E[d] = \frac{2}{3} n$$

Or if we define the area of the region as being equal to A, equation A.7 can be conveniently written as:

$$E[d] = \frac{2}{3}\sqrt{A} \qquad (A.8)^*$$

*These are the same as results derived by Larson in Reference 12.

A.3 The Time to Service an Emergency Call

The total time to service an emergency call consists of three components mentioned at the beginning of the chapter and depicted in the diagram below. These are:

(i) The time to travel to the scene of the emergency, t_1.

(ii) The time to administer treatment at the scene, t_t.

(iii) The time to travel from the scene to hospital, t_2.

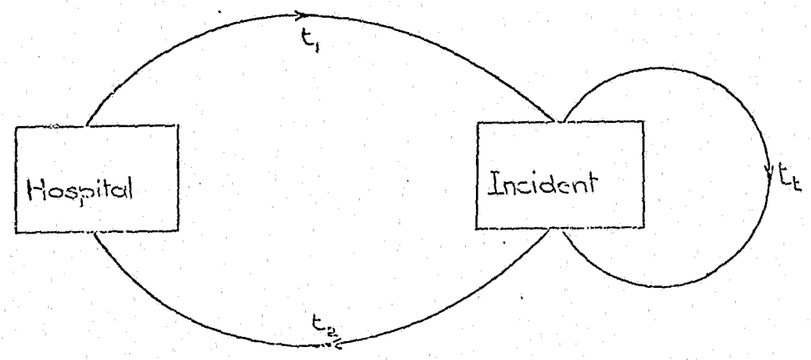


FIGURE A.3 : AMBULANCE SERVICE TIME

The total service time t_o is the sum of these three times. Whereas t_t can be assumed to be independent of the other two times, these are certainly not independent of each other. If we assume that the ambulance is located at the hospital and that traffic conditions largely determine travel speeds so that the time to travel to the scene is the same as the time to return, we can derive the distribution of the Total Travel Time, $t_r = t_1 + t_2$ from the results of the previous section.

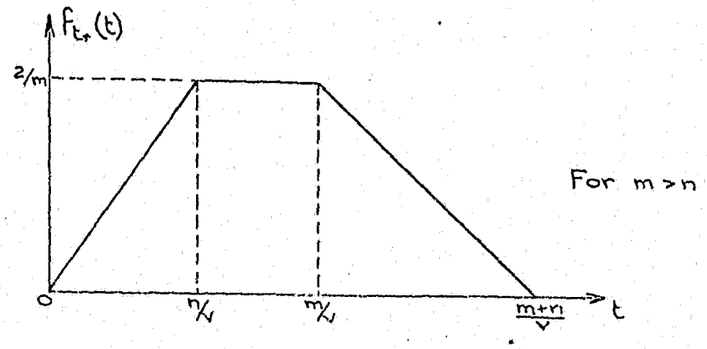
If the travel speed is v distance units per hour, then the travel time can be related to travel distance by $t_r = \frac{2d}{v}$ hours. From A.6

in (i) of Section A.2 it follows that

$$f_{t_r}(t) = \frac{4r^2 t}{mn}\left[1 - u_{-1}(y - \frac{n}{v}) - u_{-1}(t - \frac{m}{v}) + u_{-1}(t - \frac{m+n}{v})\right] \qquad (A.9)$$

$$\text{and } E[t_r] = \frac{m+n}{2v}$$

$$\text{Var.}(t_r) = \frac{m^2+n^2}{24\,v^2} \qquad (A.10)$$



The Laplace Transform of A.9 is given by:

$$f^T_{t_r}(s) = \frac{4v^2}{mns^2}\left[1 - e^{-ns/v} - e^{-ms/v} + e^{-(m+n)s/v}\right] \qquad (A.11)$$

Turning our attention to the time to administer treatment to the patient at the scene (t_t), we shall assume that it is exponentially distributed with mean $\frac{1}{\mu_t}$ hours. We have the following characterization of this distribution:

$$f_{t_t}(t) = \mu_t e^{-\mu_t t} \qquad t \geq 0$$

$$E[t_t] = \frac{1}{\mu_t}$$

$$\text{Var.}(t_t) = \frac{1}{\mu_t^2} \qquad (A.12)$$

$$\text{Laplace Transform: } f^T_{t_t}(s) = \frac{\mu_t}{\mu_t + s} \qquad (A.13)$$

Now the total service time, t_o is the sum of the independent random variables t_r and t_t. The Laplace Transform of $f_{t_o}(t)$ is given by the product of A.11 and A.13. Evaluating this product and inverting gives the desired result

$$f_{t_o}(t) = \frac{4v^2}{mn}\left[t - \frac{1}{\mu_t}(e^{-\mu_t t} - 1)\right]\left[1 - u_{-1}(t - \frac{n}{v}) - u_{-1}(t - \frac{m}{v}) + u_{-1}(t - \frac{m+n}{v})\right]$$

$$t \geq 0 \qquad (A.14)$$

Invoking our independence assumption it follows from A.10 and A.12 that

$$E[t_o] = \frac{1}{\mu_t} + \frac{m+n}{2v}$$

$$\text{Var.}(t_o) = \frac{1}{\mu_t^2} + \frac{m^2+n^2}{24\,v^2} \qquad (A.15)$$

Equation A.13 has been plotted in Figure A.4 for four different regions, assuming $\frac{1}{\mu_t}$ = 15 minutes and v = 20 m.p.h. For any reasonably large region there is a considerable deviation from the negative exponential distribution. It is therefore inappropriate to use that distribution in models of emergency ambulance services.

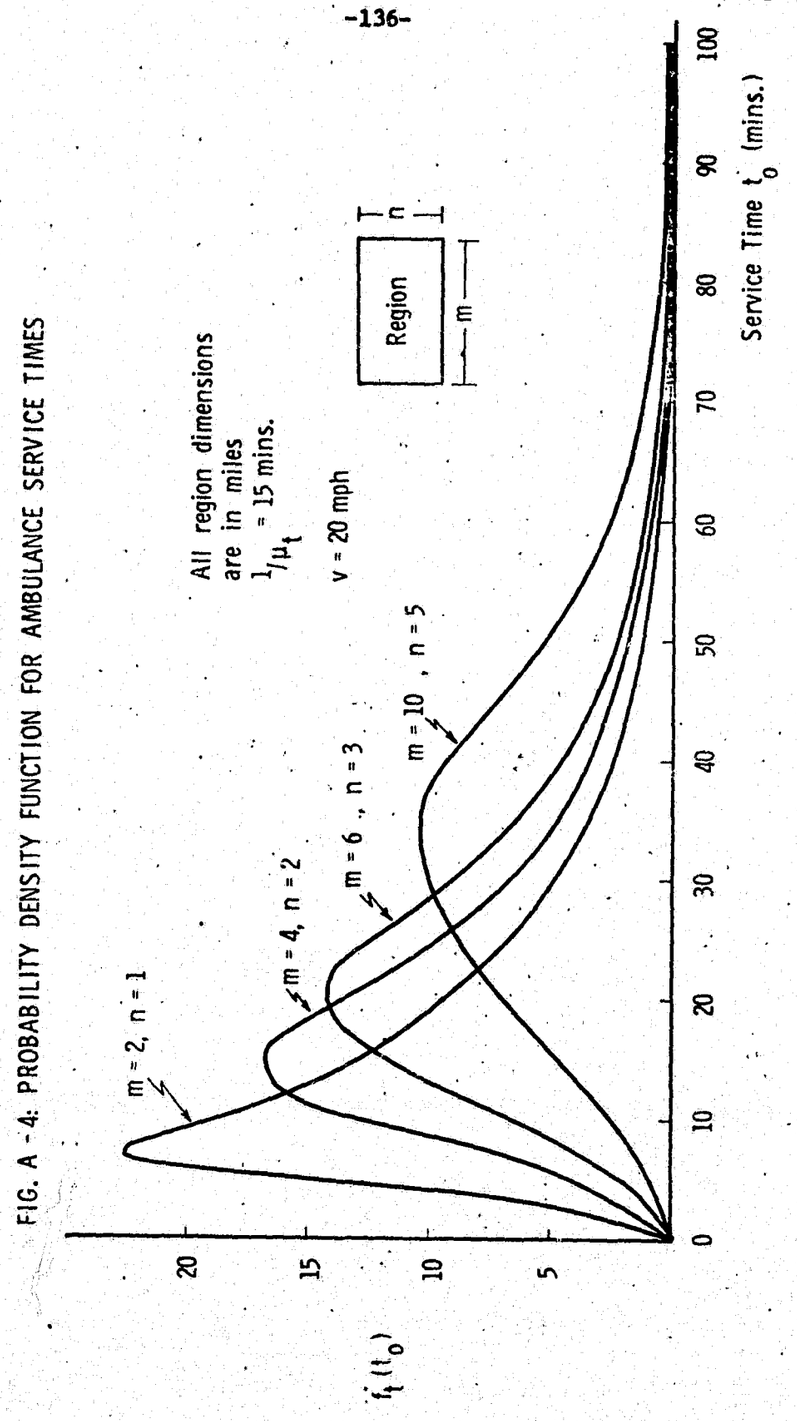


FIG. A - 4: PROBABILITY DENSITY FUNCTION FOR AMBULANCE SERVICE TIMES

A.4 Comparison with Real Systems

In New York City the emergency ambulance service is provided under conditions somewhat similar to those of our model. Selecting two very different hospital districts, St. Vincent's on Manhattan Island and King's County in Brooklyn, we may compare actual service time distributions with those predicted by the model. In both cases we assume a mean treatment time of 15 minutes. In Brooklyn the speed was assumed to be 15 m.p.h. in accordance with a study in that region[7], and in Manhattan the speed was assumed to be 10 m.p.h. Data obtained from Reference No. 3 is compared with model predictions in Figure A.5. The agreement is particularly good in the case of St. Vincent's Hospital, and the discrepancy in the case of King's County Hospital is probably due to the concentration of demand at one end of the region resulting in an unusually large travel component in the service time.

The very low mean speeds assumed in the model are not unusual for city travel. Very similar speeds are encountered by the police in Boston - see Reference No. 12, page 85. The effect of mean speed on the Expected Service Time is graphically illustrated in Figure A.6. It is clear that in congested urban areas the travel component may dominate the service time.

A result from the theory of single-server queues with Poisson arrivals and general service times is that the Expected Waiting Time is proportional to the variance of the service time.[20] We assume that a somewhat similar result holds for more than one server. Since we should like to keep the waiting time (called the dispatch delay in Chapter 4) as

FIG. A - 5(a): ST. VINCENT'S HOSPITAL, MANHATTAN

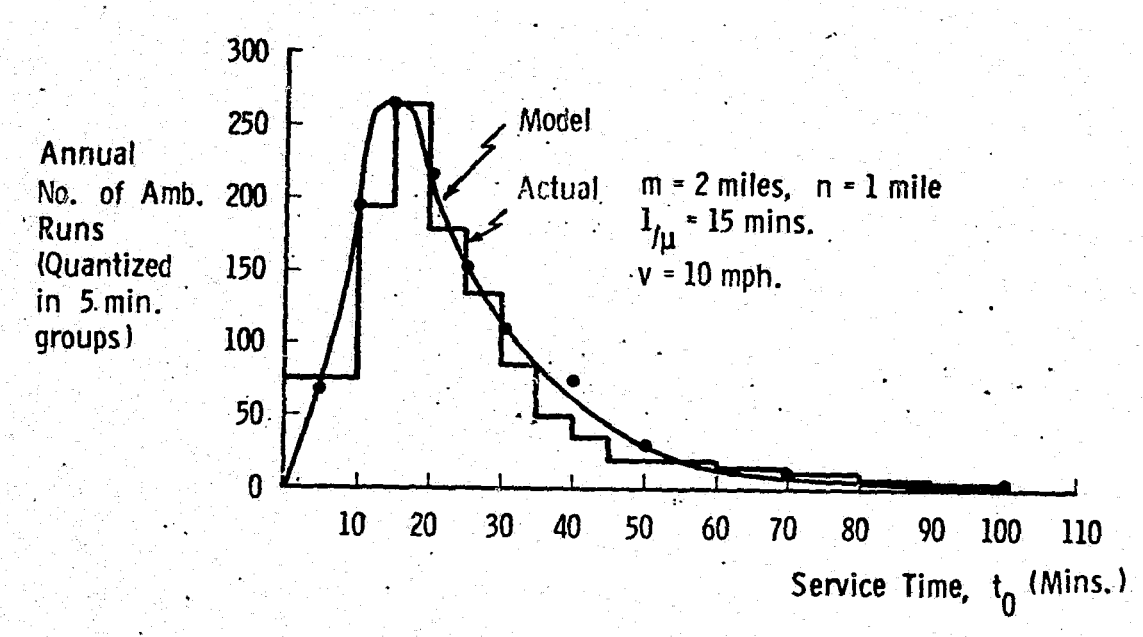
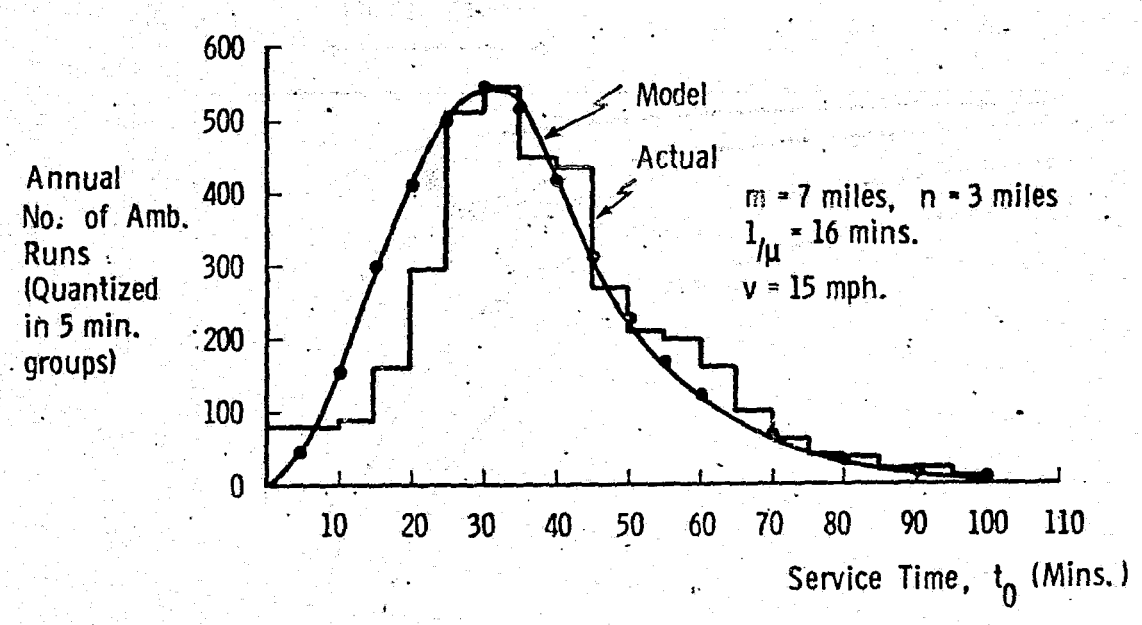


FIG. A - 5(b): KING'S COUNTY HOSPITAL, BROOKLYN



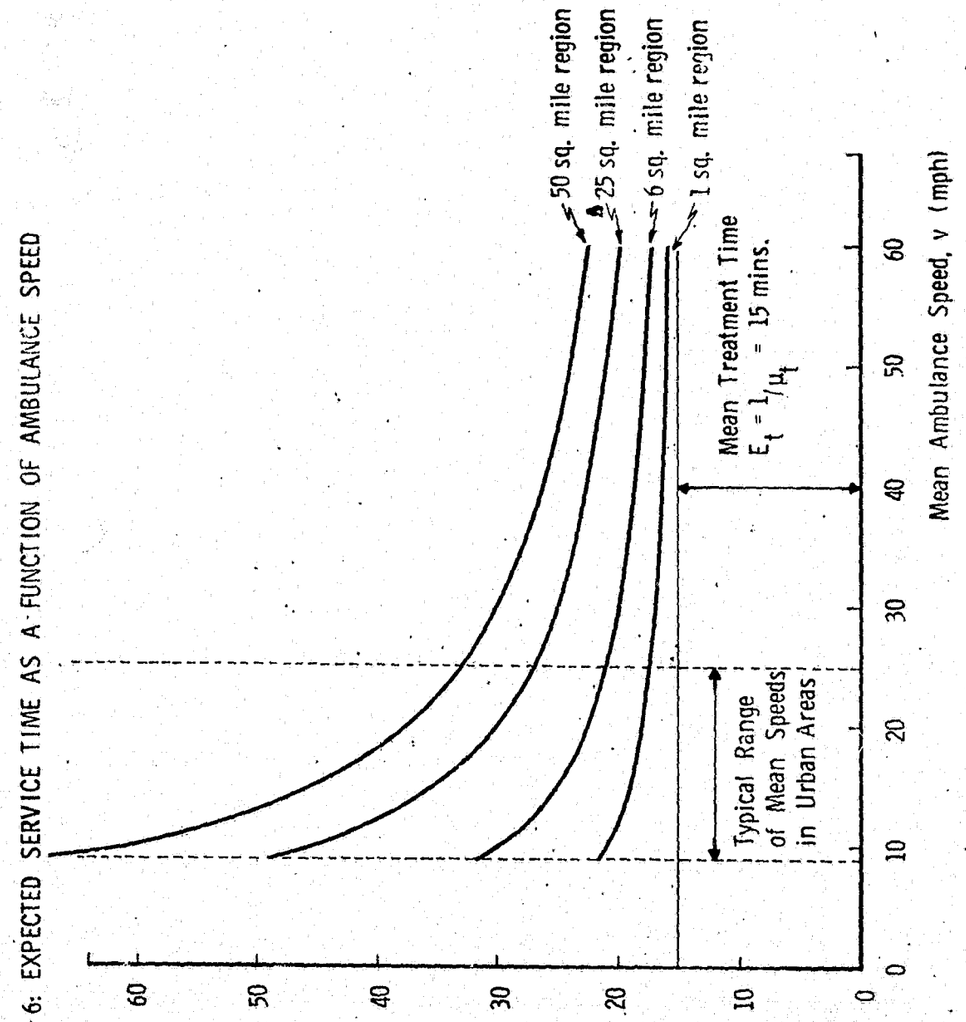


FIG. A-6: EXPECTED SERVICE TIME AS A FUNCTION OF AMBULANCE SPEED

small as possible, we should like to minimize the variance in A.14. Not surprisingly this involves increasing the speed and reducing the dimensions of the region. For any given area and speed, the square region minimizes the variance.

In summary we can observe that even with a host of simplifying assumptions the service time distribution is a somewhat involved function of region dimensions and attainable speeds. For any realistic values of size and speed the negative exponential distribution is an inadequate descriptor of the service time distribution. As the expected wait in the queuing system associated with such a service time distribution is an inverse function of the square of the speed, low speeds in urban areas threaten to have an adverse delaying effect throughout the system.

APPENDIX B: HOSPITAL DISTRICTS

B.1 Introduction

In this document we have concentrated on the events preceding the arrival of the ambulance at the scene of the emergency. We have assumed that the journey from the scene to the hospital is not particularly critical. Whereas this is true in general, there have been situations in which patients have been transported to a distant rather than a nearby hospital, with consequent adverse effects. These unnecessarily long journeys are apparently the result of habit on the part of the ambulance driver, or of a relationship existing between a few hospitals and the ambulance service.

In the next few pages we outline a simple procedure that allows us to relate an emergency at a particular location with the nearest hospital. In brief, we describe around each hospital a region such that an incident occurring inside one particular region is nearer to the hospital in that region than to any other hospital. While it is unnecessary and inappropriate to insist that <u>every</u> patient be taken to the <u>nearest</u> hospital, discouraging long and somewhat unnecessary journeys may:

(i) improve the chances for recovery of some patients;

(ii) increase ambulance availability by reducing travel times;

(iii) reduce the grossly disproportionate sharing of emergency patients among the hospitals in some towns and cities.

B.2 The Travel-Time Model

We assume as before that we have a city or town with streets laid out in the form of a rectangular grid. There are N hospitals participating in the emergency service, and we shall create N mutually exclusive and collectively exhaustive regions in the city, each associated with one hospital. Using the procedure adopted in Chapter 4 (see Figure 4.5) and Appendix A we can identify each point in the city by x,y coordinates.

The distance between an incident at point (x_i,y_i) and a hospital located at (x_4,y_4) is:

$$d = |x_H-x_i| + |y_H-y_i|$$

We can associate with each of the N hospitals a distance d from the incident at (x_i,y_i). Alternatively we can consider that subset of all the points in the city which are closer to (x_H,y_H) than to any of the remaining N-1 hospitals. We begin by examining the situation when there are only two hospitals in the city.

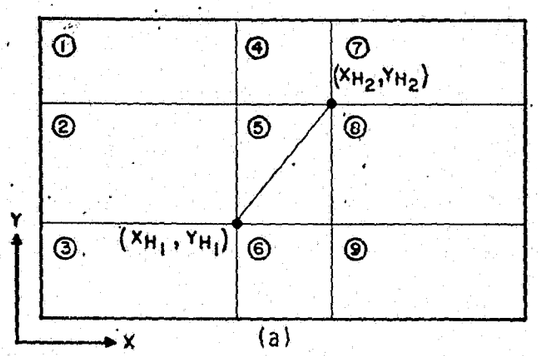
B.2.1 Two Hospitals

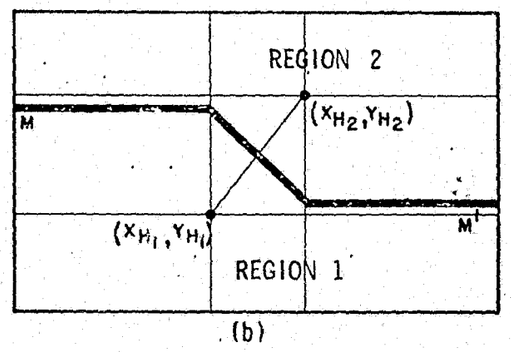
With only two hospitals in the city, there are two regions to be defined. Incidents occurring in Region 1 are closer to hospital 1 [at (x_{H1},y_{H1})] than to hospital 2 [at (x_{H2},y_{H2})]. The boundary between the regions must be defined by the equation:

$$|x_{H1}-x_i| + |y_{H1}-y_i| = |x_{H2}-x_i| + |y_{H2}-y_i| \qquad (B.1)$$

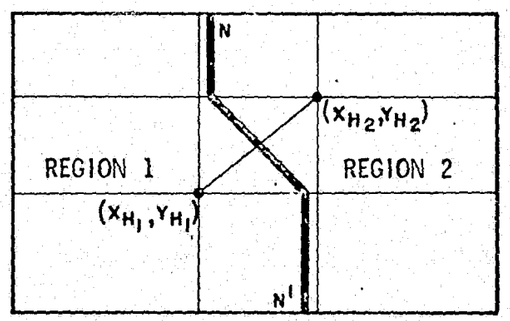
There are three possible configurations that must be considered. First, referring to Figure B.1(a), we note that:

FIG. B.1: HOSPITAL DISTRICTS WITH 2 HOSPITALS





(b)



(c)

$$x_{H2} > x_{H1}$$

$$y_{H2} > y_{H1}$$

$$(y_{H2}-y_{H1}) > (x_{H2}-x_{H1})$$

The implications of equation B.1 in areas 2, 5 and 8 are easily shown to be

In area 5: $[x_i - \frac{1}{2}(x_{H1}+x_{H2})] + [y_i - \frac{1}{2}(y_{H1}+y_{H2})] = 0$

In area 2: $y_i = \frac{1}{2}[(y_{H2}+y_{H1}) + (x_{H2}-x_{H1})]$

In area 8: $y_i = \frac{1}{2}[(y_{H2}+y_{H1}) - (x_{H2}-x_{H1})]$

These three equations define the boundary MM' shown in Figure B.1(b) between Region 1 and Region 2.

An alternative configuration of hospital location could be one so that:

$$x_{H1} < x_{H2}$$

$$y_{H1} < y_{H2}$$

$$(y_{H2}-y_{H1}) < (x_{H2}-x_{H1})$$

The solution for this situation is derived in the same way as above. and gives rise to the boundary NN' shown in Figure B.1(c). In the special case when $(y_{H2}-y_{H1}) = (x_{H2}-x_{H1})$, incidents located in areas 1 and 9 are equidistant from the two hospitals.

All possible arrangements of the two hospitals can be fitted into one of the three simple cases mentioned above. Hence it is

possible to describe around each hospital a region so that incidents in Region j are closer to Hospital j than to any other hospital (j=1,2).

B.2.2 The General Case, N Hospitals

The extension to more than two hospitals is achieved quite simply by considering the boundaries between each pair of hospitals in turn. In theory it should be necessary to consider $\frac{N(N-1)}{2}$ boundaries for the creation of N regions. In practice the number of boundaries that are candidates for inclusion in the composite boundaries is considerably less. Many can be eliminated by inspection, and it is usually only necessary to consider neighboring hospitals in constructing boundaries.

As an example we include in Figure B.2 the regional pattern resulting from the division of an area among five hospitals. Subsequent work, which is beyond the scope of this appendix, has shown that metrics other than the one used here (i.e. the sum of the X- and Y-component distances) give very similar results.

B.3 Application to the City of Boston

The police department provides the emergency ambulance service in Boston. Because of conditions that existed some years ago, a tradition has been created whereby some 60% of the emergency victims are transported to two of the city's hospitals. This practice creates conditions of serious overloading in the emergency wards of the two hospitals, while

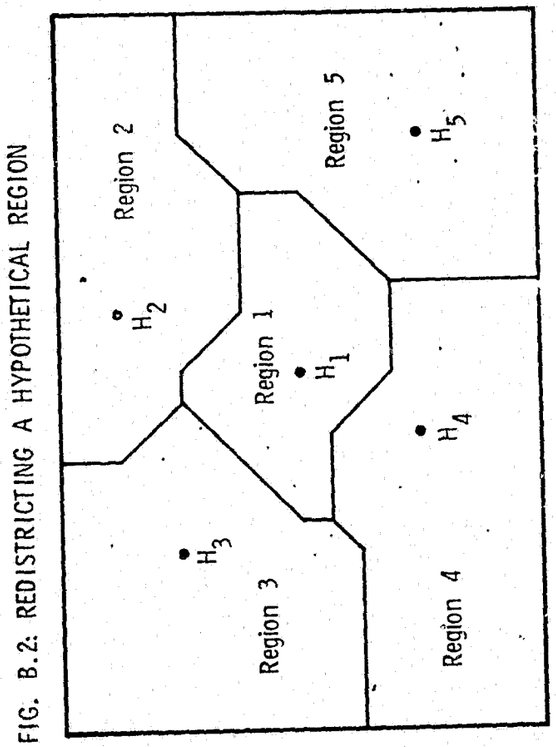


FIG. B.2. REDISTRICTING A HYPOTHETICAL REGION

other hospitals in the city have under-utilized emergency facilities. In addition patients have to undertake unnecessarily long rides in police ambulances.

Using the minimum distance criterion described above, the city can be divided into eight regions associated with the eight major hospitals in the area. Although Boston streets do not follow a rectangular grid pattern, the coordinate system from which the regions were drawn was located so that it paralleled the largest number of streets. These regions, shown in Figure B.3, are not necessarily the "best" in any sense. They do, however, represent a realistic approach to improving the quality of care received by emergency victims in the city, by reducing the time a patient spends travelling to hospital, and balancing the distribution of patients among the city's emergency wards.

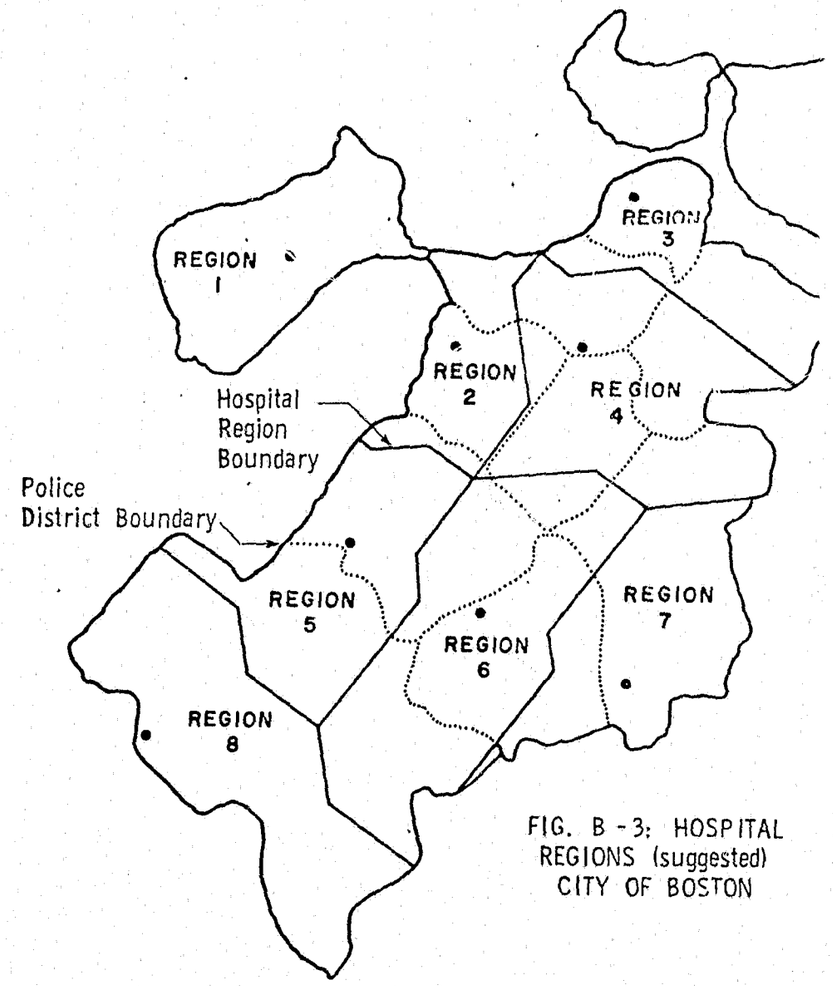


FIG. B - 3: HOSPITAL REGIONS (suggested) CITY OF BOSTON

APPENDIX C

THE COMMONWEALTH OF MASSACHUSETTS

DIVISION OF HOSPITAL FACILITIES

RULES AND REGULATIONS

RELATIVE TO

AMBULANCES

April 1968

MASSACHUSETTS DEPARTMENT OF PUBLIC HEALTH

DIVISION OF HOSPITAL FACILITIES

Rules and Regulations Relative to Ambulances

I. Definitions

A. "Ambulance" is defined as any aircraft, boat or motor vehicle, however named, whether privately or publicly owned, which is specially designed, constructed and equipped for the purpose of transporting patients. It shall not include a hearse.

B. "Patient" shall mean any individual who is sick or injured. An injured person shall include a disabled person.

C. "Person" shall include any individual, firm, partnership, association, corporation, trust, foundation, company or any group of individuals, however named, concerned with the operation of an ambulance service. "Person" shall also include any governmental agency other than the Federal Government.

D. "Ambulance Service" shall mean regularly engaging, within the Commonwealth, in the transportation by ambulance of the sick or injured.

E. A qualified attendant shall mean an individual at least 21 years of age, certified as having completed the standard and advanced American Red Cross First Aid Course or have equivalent training approved by the Division of Hospital Facilities.

F. "Commonwealth" shall mean the Commonwealth of Massachusetts.

G. "Department" shall mean the Massachusetts Department of Public Health.

H. "Division of Hospital Facilities" shall mean the Massachusetts

I. Definitions (Cont'd.)

Department of Public Health.

H. "Division of Hospital Facilities" shall mean the Massachusetts Department of Public Health, Division of Hospital Facilities.

I. "Registered Motor Vehicle" shall mean a motor vehicle currently registered by the Commonwealth of Massachusetts, Registry of Motor Vehicles.

J. "An Ambulance Certificate of Inspection" shall mean a "Certificate of Inspection" issued by the Department of Public Health to an applicant for a period of one year in accordance with the Rules and Regulations, prescribed and established under the authority of General Laws, Chapter III, 8B.

K. "Original Certificate" shall mean a certificate issued for an ambulance not previously certified, or a certificate issued for an ambulance in which there has been a change in ownership or location.

L. "Disinfection" shall mean any process, chemical or physical, by means of which pathogenic agents or disease-producing microbes are destroyed.

M. "Sanitization" shall mean a process whereby organisms present on an object are reduced in number to a level considered safe for human use.

N. "Sterilization" shall mean any process by means of which all forms of microbial life are killed.

II. Ambulance Certificate of Inspection

A. No person, either as owner, agent or otherwise, shall operate, conduct, maintain or profess by advertising or otherwise to operate

II. Ambulance Certificate of Inspection (Cont'd.)

conduct and maintain a business for transporting patients upon any way or place of the Commonwealth, unless he holds a current Ambulance Certificate of Inspection for each ambulance issued pursuant to these rules and regulations. Said Certificate of Inspection shall be framed and conspicuously posted in each ambulance and no official entry made upon an Ambulance Certificate of Inspection shall be defaced, removed or obliterated.

B. An Ambulance Certificate of Inspection shall not be required for a motor vehicle which:

1. Is rendering assistance to certified ambulances in the case of a major catastrophe or emergency.

2. Is operated from a location or headquarters outside the Commonwealth, and is transporting patients to locations within the Commonwealth. However, no such outside ambulance shall transport patients from one area to another within the Commonwealth, unless there is compliance with these rules and regulations.

C. An Ambulance Certificate of Inspection shall be issued for a specific ambulance and shall not be transferred to another ambulance.

D. Representatives of the Division of Hospital Facilities are authorized to enter and examine any ambulance, to determine if such vehicle is properly staffed, maintained and equipped in accordance with these rules and regulations. Any attempt to prevent any such representative to enter and examine any such ambulance or garage, or to examine records as required shall be punishable in accordance with "XIV Penalty" of these Rules and Regulations.

III. Application for Ambulance Certificate of Inspection

A. Applications for an original or renewal Ambulance Certificate of Inspection shall be made in writing upon forms provided by the Department and shall contain:

1. Name and address of the applicant and of the owner of the ambulance.
2. Trade or other name, if any, under which the applicant does business and/or proposes to do business.
3. Training and experience of the applicant and attendants in the transportation and care of patients.
4. Description of each ambulance, including the make, model, year of manufacture; vehicle identification number; current Massachusetts registration number; and the color, insignia, name, monogram or other distinguishing characteristics, if any, to be used to designate applicant's ambulance.
5. Location of the place or places from which it is intended to operate.
6. Such information as is necessary for the Department of Public Health to carry out its responsibilities under these rules and regulations.

IV. Change in Ownership

A. Upon change of ownership, an Ambulance Certificate of Inspection shall terminate and the new owner shall be required to file an application for an Ambulance Certificate of Inspection, in conformance with all the requirements for an original or renewal Ambulance Certificate of Inspection, herein set forth.

IV. Change in Ownership (Cont'd.)

B. When a registered or certified notice is received by mail by the Division of Hospital Facilities that a change of ownership of an ambulance has occurred or a new ambulance has been acquired, such notice shall have the effect of an ambulance certificate of inspection for a period not to exceed thirty days.

V. Certification Procedure

A. The Director of the Division of Hospital Facilities upon receipt of an application for an Ambulance Certificate of Inspection shall cause to be inspected, the ambulance, equipment and premises designated in each application hereunder to determine compliance with the rules and regulations.

B. The Department shall issue an Ambulance Certificate of Inspection for a specified ambulance to be valid for a period of one year.

VI. Revocation of Ambulance Certificate of Inspection

A. The Department may revoke an Ambulance Certificate of Inspection issued hereunder for cause. Failure of a holder of an Ambulance Certificate of Inspection to comply with the rules and regulations of the Department promulgated hereunder, shall be sufficient cause for revocation, after a public hearing held in accordance with General Laws, Chapter 30A.

VII. Return of Ambulance Certificate of Inspection

A. Each Ambulance Certificate of Inspection shall be returned to the Division of Hospital Facilities immediately by registered or

VII. Return of Ambulance Certificate of Inspection (Cont'd.)

certified mail upon:

1. Expiration of certificate.
2. Revocation of certificate.
3. Change in ownership of ambulance.
4. Change of name of ambulance service.
5. Discontinuance of use of vehicle as an ambulance

VIII. Ambulance Equipment

A. Each ambulance and its equipment shall be maintained in a sanitary manner and in good operating condition.

B. Each ambulance shall be equipped with the following:

1. Two-way radio communication system.
2. Recording Tachometer.
3. Siren.
4. Flashing Red roof light.
5. Fire extinguisher (Underwriter's Laboratory) approved.
6. Explosion proof flashlight.

C. Each ambulance shall be equipped with the following equipment or its equivalent when approved by the Division of Hospital Facilities:

1. Hinged half-ring lower extremity split with web straps for ankle hitch.
2. Two or more padded boards 4 1/2 feet long and 3 inches wide, and two or more similar padded boards 3 feet long by 3 inches wide, of material comparable to four-ply wood, for coaptation splinting of fracture of leg or thigh.

VIII. Ambulance Equipment (Cont'd.)

3. Two or more padded 15 inch by 3 inch wood or cardboard splints for fractures of the forearm.
4. Short and long spine boards with accessories.
5. Oxygen tanks with regulators and single use disposable masks of assorted sizes.
6. Hand-operated bag-mask resuscitation unit with adult, child and infant size masks, capable of being attached to oxygen supply.
7. Simple suction apparatus with catheter.
8. Mouth to mouth, two-way resuscitation airways for adults and children.
9. Oropharyngeal airways.
10. Mouth gags.
11. Universal dressing
12. Sterile gauze pads.
13 1, 2 and 3 inch adhesive tape.
14. Six inch by 5 yard soft roller type bandages.
15. Triangular bandages.
16. Safety pins, large size.
17. Bandage shears.
18. Collapsible stretcher with straps.
19. Two sandbags.

D. Linen and Patient Equipment

1. At least two pillows with removable washable protective covers.
2. A minimum of six individually packaged pillow cases, preferably disposable.

VIII. Ambulance Equipment (Cont'd.)

3. A minimum of six individually packaged sheets, preferably disposable.
4. Sufficient towels to protect the patient's head and face, as indicated.
5. A sufficient number of washable blankets in accordance with seasonal requirements.
6. A sufficient supply of laundry bags, preferably disposable, including a special color or suitable identification for precaution linen.
7. A sufficient supply of towels, tissues and paper bags.
8. Emesis basins, preferably disposable.
9. Sanitized wrapped bed-pan.
10. Sanitized wrapped urinal.

E. Storage Facilities

1. There shall be adequate storage facilities for:
   a. Clean supplies and equipment.
   b. Clean linen.
   c. Soiled linen.
   d. Waste.

IX. Standards for Ambulance Certification

A. Each ambulance shall contain equipment conforming with the rules and regulations provided for herein and shall be:
1. Staffed at all times with a minimum of two qualified attendants, one of whom may be the driver.

IX. Standards for Ambulance Certification (Cont'd.)

2. Maintained in such a manner as to insure the health and safety of the patient.
3. Used exclusively for the purpose of transporting sick, injured or disabled persons.

B. When not in service, an ambulance shall be protected in such a manner that it does not become unsanitary.

X. Records

A. Records of service rendered shall be maintained and stored in a satisfactory manner for a minimum of two years.

B. Record content shall include the following:
1. Date and time of arrival to transport patient, time of arrival at destination.
2. Name, address, age and sex of patient.
3. Transported: From - - To - -
4. Name of attendants.
5. First aid administered with date, time and signature of person administering same.

C. Personnel Records
1. A personnel file shall be maintained for attendants and shall include their qualifications and training.

D. Records required herein shall be available for inspection by representatives of the Division of Hospital Facilities.

VIII. Ambulance Equipment (Cont'd.)

3. A minimum of six individually packaged sheets, preferably disposable.
4. Sufficient towels to protect the patient's head and face, as indicated.
5. A sufficient number of washable blankets in accordance with seasonal requirements.
6. A sufficient supply of laundry bags, preferably disposable, including a special color or suitable identification for precaution linen.
7. A sufficient supply of towels, tissues and paper bags.
8. Emesis basins, preferably disposable.
9. Sanitized wrapped bed-pan.
10. Sanitized wrapped urinal.

E. Storage Facilities

1. There shall be adequate storage facilities for:
   a. Clean supplies and equipment.
   b. Clean linen.
   c. Soiled linen.
   d. Waste.

IX. Standards for Ambulance Certification

A. Each ambulance shall contain equipment conforming with the rules and regulations provided for herein and shall be:

1. Staffed at all times with a minimum of two qualified attendants, one of whom may be the driver.

XIV. Penalty (Cont'd.)

Certificate of Inspection, as set forth in these Rules and Regulations, or whoever being a holder of an Ambulance Certificate of Inspection as required herein, violates any provisic of these Rules and Regulations as established by the Department in accordance with General Laws, Chapter III, Section 8B, shall be punished by a fine of not more than $500.00 for any particular offense. A separate and distinct offense shall be deemed to have been committed on every day during which the violation continues after written notice thereof by the Department to the authority to whom the Ambulance Certificate of Inspection was issued.

XV. Separability

If any section, subsection, sentence, clause, phrase or portion of these Rules and Regulations is for any reason held invalid or unconstitutional by any court of competent jurisdiction, such portion shall be deemed a separate, distinct and independent provision and such holding shall not affect the validity of the remaining portions hereof.

XVI. An Emergency Situation

These rules and regulations shall not preclude the reasonable omission of any of the foregoing requirements when a law enforcement officer or a representative of a fire department determines an emergency exists.

XI. Personnel

A. Ambulance Attendants

1. The ambulance attendants shall be well groomed, appropriately attired in uniform, and shall practice good personal hygiene, including handwashing.

XII. Sterilization, Sanitization and Disinfection

A. Equipment and supplies requiring sterilization shall be processed by methods approved by the Division of Hospital Facilities.

B. Written policies for the routine disinfection of the interior of the ambulance, and for the disinfection and sanitization of all equipment, unless disposable, shall be established and enforced.

C. All equipment, unless disposable, shall be properly cleansed and sanitized after each use.

D. When an ambulance has been used to transport a patient known to have a communicable disease, the interior of the ambulance shall be disinfected before the next patient is transported.

XIII. Linen

A. Linen, unless disposable, shall be laundered after each use.

B. Disposable linens shall be discarded after each use.

C. Soiled linen, including disposable linen, shall be handled in such a manner as to avoid contamination of equipment and personnel.

XIV. Penalty

Whoever advertises, announces, establishes or maintains an ambulance and/or ambulance service, as defined herein, without the required Ambulance

13. Larson, Richard C., Models for the Allocation of Urban Police Patrol Forces. Technical Report No. 44. M.I.T. Operations Research Center, Cambridge, Mass. Nov. 1969.

14. Mancgold, Silver, Owen, and Sigmond, "An Overview of Emergency Medical Services." Journal of the American Medical Association. Vol. 200, No. 4. April, 1967.

15. Owen, Joseph K., "Emergency Services Must be Reorganized." The Modern Hospital, Vol. 107, No. 6, Dec., 1966.

16. Pocket Data Book: U.S.A. 1969. Bureau of the Census, U.S. Dept. of Commerce.

17. Public Health Service, U.S. Dept. of Health, Education and Welfare, Availability of Federal Funds for the Operation of Emergency Ambulance Services. March, 1969.

18. Rutstein, David D., The Coming Revolution in Medicine. M.I.T. Press. 1967.

19. Sevastyanov, B. A., "An Ergodic Theorem for Markov Processes and Its Application to Telephone Systems with Refusals." Theory of Probability and Its Applications. Vol. II, No. 1. 1957.

20. Syski, R., Introduction to Congestion Theory in Telephone Systems. Oliver and Boyd, London. 1960.

21. Savas, E. S., "Simulation and Cost Effectiveness Analysis of New York's Emergency Ambulance Service." Management Science, 15, 608. 1969.

22. Stevenson, K. A., Emergency Ambulance Services in U.S. Cities - An Overview. Technical Report No. 40. M.I.T. Operations Research Center, Cambridge, Mass. August, 1968.

23. Weinerman, E. R., "Yale Studies in Ambulatary Care." New England Journal of Medicine. Vol. 272, No. 18. 1965.

24. West, Kleinman, Taylor, Majors and Mitchell, Study of Emergency Ambulance Operations - A Preliminary Report. May, 1964.

25. Bellman, R. E. and S. E. Dreyfus, Applied Dynamic Programming. Princeton University Press, Princeton, New Jersey. 1962.

REFERENCES

1. Accidental Death and Disability: The Neglected Disease of Modern Society. National Academy of Sciences and the National Research Council. September, 1966.

2. Cooper et al., Description and Analysis of Eighteen Proven Emergency Ambulance Service Systems. National Association of Counties Research Foundation. Prepared for the U.S. Dept. of Transportation. 1968.

3. Dimendberg, D. C., An Analysis of the Emergency Ambulance Service of the Dept. of Hospitals for August, 1966. Dept. of Hospitals, City of New York.

4. Drake, Alvin W., Fundamentals of Applied Probability Theory. McGraw-Hill, New York. 1967.

5. Dunlap and Associates, Inc., Economics of Highway Emergency Ambulance Services. Prepared for the U.S. Dept. of Transportation. July, 1968.

6. Godlund, Sven, "Population, Regional Hospitals, Transport Facilities, and Regions." Dept. of Geography, Royal University of Lund, Sweden. 1961.

7. Gordon, G. and K. Zelin, A Simulation Study of Emergency Ambulance Service in New York City. I.B.M. Technical Report No. 320-2935. March, 1968.

8. Highway Safety Program Manual - Vol. 11: Emergency Medical Services. U. S. Dept. of Transportation: National Highway Safety Bureau. January, 1969.

9. Hospital Services in the U.S.S.R. Report of the U.S. delegation on hospital systems planning, June 26 - July 16, 1965. Public Health Service Publication No. 930-F-10.

10. Jackson, Laura G., Hospital and Community. Macmillan, New York. 1964.

11. Jacobs, A. R. and C. P. McLaughlin, "Analyzing the Role of the Helicopter in the Emergency Medical Care for a Community." Medical Care, Vol. 5, No. 5. Sept. - Oct., 1967.

12. Larson, Richard C., Operational Study of the Police Response System. Technical Report No. 26. M.I.T. Operations Research Center, Cambridge, Mass. Dec., 1967.

END